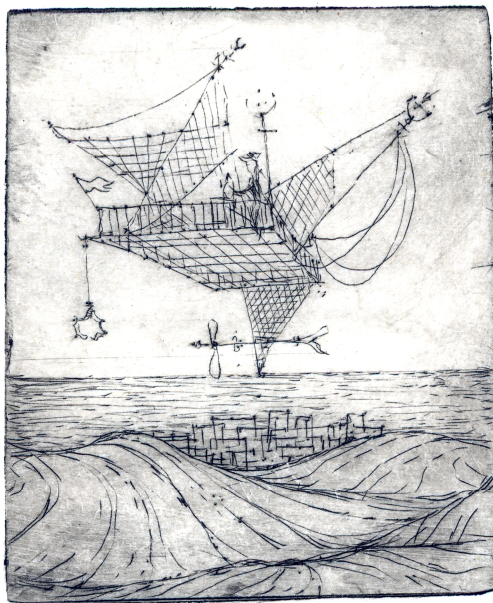
Manuel Barbera

Linguistica dei corpora e linguistica dei corpora italiana. Un’introduzione.

Milano, Qu.A.S.A.R. s.r.l.
2013

ISBN-10: 88-87193-28-2
ISBN-13: 978-88-87193-28-2

La versione e-book è scaricabile gratuitamente da http://www.bmanuel.org/

Ma chi non può quel che vuole, quel che può voglia.

Michel Barbi, *Studi sul canzoniere di Dante*, Firenze, 1915, capo VIII.

0. Introduzione*.

Quello che qui si presenta non è un manuale tecnico di statistica testuale pensato per l'ingegnere computazionale, che forse non ne avrebbe neppure bisogno, sia perché è di solito interessato ad operazioni più complesse della "semplice" linguistica dei corpora, sia perché la sua bibliografia di riferimento c'è già, sia pure in lingua inglese. La mia intenzione è invece di rivolgermi al linguista (ancorando chiaramente la linguistica dei corpora alla storia della linguistica, sfatando la credenza diffusa tra i linguisti generali che questa sia solo roba da praticoni) ed a tutti gli altri potenziali utenti di corpora non linguisti, *in primis* filologi e storici della lingua, ma non solo, offrendo loro una presentazione volutamente molto sintetica, quasi un breviario laico. Iniziale e non iniziatico.

Il *côté* informatico e quello statistico non saranno quindi prioritari (tanto poi se son rose fioriranno, ma prima *sinite parvulos ad me*), quello che conta è prima una mossa culturale definita, e poi mettere praticamente in grado chiunque di tracciarsi la propria strada. La bibliografia al fondo serve appunto a ciò: permettere al lettore, una volta che abbia trovato il proprio orientamento, di andare oltre nella direzione che deciderà.

Ed a questo secondo scopo presenterò per il solo italiano un panorama delle risorse liberamente disponibili (ad esclusione di quelle commerciali), dopo essermi soffermato sugli elementi, prima, teorico-storici e, poi, costitutivi del corpus, sui quali mi soffermerò di più, avendo in mente tanto l'utente ingenuo quanto il potenziale costruttore di corpora fai-da-te, che di solito non è un tecnico.

Ontologie ed annotazioni semantiche, tecniche statistiche avanzate (corpora e dati testuali), acustica (sintesi e riconoscimento vocale) e filologia elettronica sono discipline autonome, che, certo, qual più qual meno, si intersecano con la linguistica dei corpora, ma richiederebbero, per essere adeguatamente affrontate, trattazioni separate (che

---

* Ringrazio, a vario titolo, Giorgio Graffi per avermi dato l'idea di scrivere questo testo, e Ludwig Fesenmeier, Carla Marello, Stefano Ondelli ed Andrea Villarini per i preziosi suggerimenti.

sono certo auspicabili); qui se ne faranno solo pochi cenni, quando indispensabile.

Per facilitare la fruizione del manuale anche ai “catecumeni” si è cercato di non dare troppo per presupposto e di salvaguardarne la facilità di lettura contenendo al minimo le note ed i riferimenti bibliografici nel testo, avendo invece cura di offrire una bibliografia analitica e commentata in appendice.

La pubblicazione, infine, del volume sotto forma di e-book gratuito, rilasciato sotto licenza Creative Commons Share Alike, obbedisce ad un preciso programma di diffusione e circolazione della cultura (in opposizione all’attentato istuzionale che ne sta venendo perpetrato) propugnato anche in più parti del libro.

0.1 Cos’è in breve la linguistica dei corpora. «Un corpus è una collezione di testi selezionati e organizzati per facilitare le analisi linguistiche» recita la Wikipedia italiana. Non è proprio vero (cfr. la definizione formale che daremo nel § 2.1), ma da qualche parte, in effetti, bisogna pur partire, se da qualche parte si vuole andare. Per dirla con Franco Crevatin «il problema resta di capire oggi quello che vorremo trovare domani: come facciamo ad andare dove vogliamo andare? – per ricordare l’immortale Totò».

Come definizione operazionale (cioè da usare giusto come provvisorio trampolino di partenza), però, forse è meglio una definizione della *linguistica dei corpora*, più che dei *corpora* (di cui, più o meno, abbiamo già qualche nozione intuitiva approssimativa). La definizione, allora, sarebbe ancora più lassa e, peggio, circolare, ma, almeno, abbastanza ecumenica: “la *linguistica dei corpora* è quel tipo di linguistica che usa come suoi strumenti principali i corpora”.

Come precisare questa intuizione iniziale è quello che vedremo nelle prossime pagine.

0.2 Anglicismi e linguistica dei corpora: un’avvertenza preliminare. Una necessaria, preliminare, avvertenza, che va anche a

confermare quanto si dirà a proposito del radicamento della nostra disciplina nella tradizione grammaticografica italiana e della sua intrinseca maturazione, concerne il trattamento dei numerosi anglicismi tecnici che vi sono invalsi, che sono stati ripetutamente oggetto di studio e normalizzazione da parte di Carla Marello e mia.

La tradizione cui bisogna in questo caso rifarsi è soprattutto quella dell'antipurismo pragmatico e moderato (ma già il purismo italiano della Crusca ha caratteristiche speciali: cfr. oltre § 1.3) che ha il suo più alto corifeo in Leopardi. V'è un passo dello *Zibaldone* che detta chiaramente la via e che giova rileggere, idealmente sostituendo al francese l'inglese, ed alla lingua filosofica quella scientifica:

> Per li nostri pedanti il prendere noi dal francese o dallo spagnuolo voci o frasi utili e necessarie, non è giustificato dall'esempio de' latini *classici* che altrettanto faceano dal greco, come Cicerone massimamente e Lucrezio, né dall'autorità di questi due e di Orazio nella Poetica, che espressamente difendono e lodano il farlo. [...] Ben è vero che la greca letteratura e [3193] filosofia fu, non sorella, ma propria madre della letteratura e filosofia latina. Altrettanto però deve accadere alla filosofia italiana, e a quelle parti dell'italiana letteratura che dalla filosofia devono dipendere e da essa attingere, per rispetto alla letteratura e filosofia francese. La quale dev'esser madre della nostra, perocché noi non l'abbiamo del proprio, stante la singolare inerzia d'Italia nel secolo in che le altre nazioni d'Europa sono state e sono più attive che in alcun'altra. E voler creare di nuovo e di pianta la filosofia, e quella parte di letteratura che affatto ci manca (ch'è la letteratura propriamente moderna) [...] sarebbe cosa, non solo inutile, ma stolta e dannosa, mettersi a bella posta lunghissimo tratto addietro degli [3194] altri in una medesima carriera, volersi collocare sul luogo delle mosse quando gli altri sono già corsi tanto spazio verso la meta, ricominciare quello che gli altri stanno perfezionando; e sarebbe anche possibile, perché né i nazionali né i forestieri c'intenderebbono se volessimo trattare in modo

> affatto nuovo le cose a tutte già note e familiari, e noi non ci cureremmo di noi stessi, e lasceremmo l'opera, vedendo nelle nostre mani bambina e schizzata, quella che nelle altrui è universalmente matura e colorita; e questo vano rinnovamento piuttosto ritarderebbe e impaccerebbe di quel che accelerasse e favorisse gli avanzamenti della filosofia, e letteratura moderna filosofica. [...] se vuol dunque l'Italia avere una filosofia ed una letteratura moderna filosofica, le quali finora non ebbe mai, le conviene di fuori pigliarle, non crearle da se [*sic*]; e di fuori pigliandole, le verranno principalmente dalla Francia (ond'elle si sono sparse anche nelle altre nazioni [...]), e vestite di modi, forme, frasi e parole francesi (da tutta l'Europa universalmente accettate, e da buon tempo usate): dalla Francia, dico, le verrà la filosofia e la moderna letteratura, come altrove ho ragionato; e volendole ricevere, nol potrà altrimenti che ricevendo altresì assai parole e frasi di là, ad esse intimamente e indivisibilmente spettanti e fatte proprie; [3196] siccome appunto convenne fare ai latini delle voci e frasi greche ricevendo la greca letteratura e filosofia; e il fecero senza esitare.

Riappropriarci della tradizione che ci è propria (come qui si farà: cfr. § 1) non significa sbarazzarsi di quello che altri hanno già elaborato, e poi doverlo "reinventare". In pratica, la strategia che Barbera e Marello avevano abbozzato fin dal 2003 si basava su una certa generosità ad ammettere l'uso di termini di origine straniera ritenuti tecnicamente "indispensabili" (per specificità e/o diffusione internazionale), e sulla accettazione del loro ingresso, almeno iniziale, nella lingua come prestiti non adattati.

Questo orientamento, fattualmente, si traduce nella considerazione di alcuni fattori da tenere in conto per decidere quello che sia da considerarsi "prestito" e non voce *tout court* straniera:

(1) la presenza *de facto* di una voce di origine straniera in un lessico specialistico;

(2) il suo uso e frequenza anche fuori dal singolo dominio specialistico di partenza
  (a) nella lingua parlata usuale,
  (b) in più domini specialistici;

(3) la presenza di derivati a morfologia italiana e la loro diffusione
  (a) in condizioni del tipo (2),
  (b) in condizioni del tipo (1);

(4) la diffusione internazionale del prestito.

La decisione, giocoforza, sarà parametrica e le "condizioni" sopra elencate vanno applicate "a catena": la semplice presenza in un lessico specialistico (1) non basta, infatti, da sola a far accettare un prestito, ma già la soddisfazione della condizione (2) può da sola rendere il prestito accettabile, soprattutto se (b) fosse presente in più campi specialistici, e meglio ancora se fosse soddifatta anche la terza condizione (3), poiché l'accettabilità di un prestito è tanto più alta quanto più alta è la frequenza della base e dei suoi derivati e soprattutto la potenziale diffusione dei derivati fuori del dominio specialistico di partenza: la creazione di una "famiglia lessicale" prova in sé l'acclimatamento della base straniera nel lessico ospite. Nei casi più incerti, infine, sarà il fattore internazionale (4) a far pendere l'ago della bilancia da una parte o dall'altra.

Le conseguenze normo-tipografiche di ciò, per evitare ad un testo stampato vuoi il ridicolo di plurali come *films*, vuoi la eccessiva pesantezza dei troppi corsivi, sono:

(a) i prestiti accettati vanno in tondo e non in corsivo in quanto parole non più straniere
(quindi: "file" e "corpus", e non "*file*" e "*corpus*");

(b) quanto alla formazione del plurale,
    (1) i prestiti da lingue moderne rimangono invariati
    (quindi: "i file" e non "i *files*")
    (2) i prestiti da lingue classiche (mediati o meno dall'inglese; e lo stesso vale per il tedesco, tra le lingue moderne forse la più "classica": tanto sono da amare *i Lieder* quanto da aborrire **i lied*) sono pluralizzati come da grammatica (quindi: "i corpora" e non "i corpus", nonostante siano ormai abbastanza diffusi anche i plurali invariati, e talvolta questo troppo disinvolto comportamento è stato persino accettato da qualche lessicografo)

(c) la derivazione avviene secondo le normali regole italiane: prestiti non adattati in derivazione producono prestiti adattati
(quindi: "tag" > "taggare" > "taggato").

(d) la ortografia originale viene tendenzialmente mantenuta in quanto distintiva anche delle famiglie derivazionali
(quindi: *token* > "tokenizzato")

(e) le forme con trattino o spazio nell'originale se possibile sono univerbate con caduta del trattino o dello spazio
(quindi: *mark-up* e *home page* > "markup" e "homepage"; caso diverso però è quello di POS-taggato ecc., in quanto POS è una sigla mantenuta come tale in maiuscolo).

Lo scopo, naturalmente, di queste "norme" è sì quello di fornire un criterio, in primo luogo normo-tipografico, ma anche quello di preservare una prospettiva "internazionalistica" in cui inserire il fenomeno.

1. La linguistica dei corpora nella storia della linguistica: tradizione anglofona *vs* italiana. La collocazione della linguistica dei corpora nella storia della linguistica occidentale, ed il suo

confronto con la grammatica generativa[1] (perché a tale la questione spesso è stata ridotta), è cruciale per la definizione della materia, e per le sue prospettive future.

1.1 La nascita della linguistica dei corpora. La storia "inglese" che normalmente si racconta è che il capostipite di tutti i corpora attuali è il *Brown Corpus of American Written English*, compilato da Winthrop Nelson Francis ed Henry Kučera alla Brown University del Rhode Island e pubblicato nel 1964. E questo è certo il primo corpus a soddisfare in tutto e per tutto la moderna definizione formale qui data nel § 2.1. Inoltre si aggiunge di solito che chi realmente ha inaugurato tale tradizione fu Charles Carpenter Fries, quando negli anni '50 (era già anziano: nacque nel 1887), prima dunque della grande era dei computer, pubblicò una grammatica descrittiva della lingua inglese parlata basandosi sulla registrazione di 250.000 parole di conversazioni telefoniche.

Anche se la tradizione anglofona, nata da tanti lombi, è certo diventata la più rilevante nel panorama mondiale (tanto da dettar legge fin nella terminologia, cfr. qui § 0.2), anche quella italiana, di solito taciuta nella manualistica (prevalentemente di origine inglese) non è molto da meno. Se l'America può vantare, al confine tra l'epoca degli avi e quella dei padri, un Fries, noi dovremmo adeguatamente valorizzare l'opera del padre Roberto Busa SJ su Tommaso d'Aquino, iniziata nel 1949 ma comunque già fondata su spogli elettronici: pare, anzi, che sia proprio Busa a dover essere considerato il vero capostipite della nostra *gens*. Capostipite (classe del '13), peraltro, fino a pochi anni fa ancora ben presente ed attivo: se l'incontro del padre con

---

[1] L'unica definizione, a quel che mi consta, che Chomsky, il suo fondatore, ne abbia mai dato è quella contenuta in nota nella sua *Linguistica Cartesiana*: «by a "generative grammar" I mean a description of the tacit competence of the speaker-hearer that underlies his actual performance in production and perception (understanding) of speech. A generative grammar, ideally, specifi es a pairing of phonetic and semantic representations over an infi nite range; it thus constitutes a hypothesis as to how the speaker-hearer interprets utterances, abstracting away from many factors that interweave with tacitcompetence to determine actual performance».

Watson all'IBM di New York nel 1949 fa ormai parte dell'epopea, così come il suo primo *Saggio* del 1951, la versione online del suo fondamentale *Index Thomisticus* è infatti opera del nuovo millennio.

1.2 Antigenerativismo e tradizione anglofona. La linguistica dei corpora anglosassone si è di solito voluta presentare come una radicale novità, accentuando gli aspetti quantitativi sui qualitativi, e contrapponendosi, a volte in modo esasperato, al generativismo come roccaforte empiristica, perlopiù in modo assai generico (come nella manualistica più diffusa, quale il classico manuale di Tony McEnery ed Andrew Wilson del 1996 e riedito nel 2001; significativamente la questione è stata però assai ridimensionata nel recente manuale di Tony McEnery ed Andrew Hardie) e più raramente in modo meditato e filosoficamente consapevole (mossa propria quasi solo di Geoffrey Sampson); così l'enfasi è vertita sul ricorso esclusivo ai dati presenti nei corpora, spesso ipostatizzati come soli oggetti linguistici possibili (il cosiddetto procedimento *corpus driven*) in palese ostilità all'introspezione propugnata dal paradigma generativo.

In ambienti anglofoni, si è quindi assistito ad una vera costruzione della linguistica dei corpora come una sorta di antigenerativismo radicale: reazione non incomprensibile se le opinioni drasticamente espresse da Chomsky nel '58 ad un autorevole convegno in Texas («Any[2] natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list») hanno effettivamente determinato il blocco pressoché completo dei finanziamenti ai progetti computazionali di tutta una generazione. Per usare un noto slogan, viene da chiedersi: perché tanto odio?

La risposta risiede probabilmente nel ruolo chiave giocato dalla polemica antibehaviourista nella creazione della teoria generativa,

---

[2] «Tutti i corpora saranno distorti. Alcuni frasi non ci saranno perché sono ovvie, altre perché sono false, altre ancora perché sono scortesi. Il corpus, se naturale, sarà così brutalmente distorto, che la sua descrizione non sarà più che una semplice lista».

tanto che echi di quella *querelle* continuano a risuonare anche quando il mondo della ricerca è ormai radicalmente cambiato. La parabola linguistica di Chomsky, infatti, si è aperta anche, a due soli anni dalle epocali *Syntactic Structures* che segnano la nascita ufficiale del programma generativo, con una veemente (e storicamente mortale) recensione-stroncatura del behaviourismo[3], impersonato in un lavoro di Skinner, il più eminente dei behaviouristi: come se, appunto, fosse proprio il behaviourismo estremo il primo vero nemico con cui la nascente teoria generativa dovesse fare i conti.

E non solo Fries era un dichiarato behaviourista, ma i legami iniziali tra linguistica dei corpora e behaviourismo sono evidenti, non fosse che perché essi rappresentano forme diverse di un radicale esternismo[4]. Molte delle polemiche tra linguistica empirica (per usare l'ottima etichetta di Sampson) e linguistica generativa riproducono in parte quei vecchi schemi, ed avvengono in realtà solo tra le ali più oltranziste dei due schieramenti, soprattutto dove la matrice behaviourista o "empirista" è più radicata, come negli *States* ed in particolare in Gran Bretagna.

---

[3] Il *comportamentismo* (o *behaviourismo*) è stato il più importante tentativo di fondare la psicologia su basi empiriste. Fondata dall'americano John Broadus Watson (1878-1958) nel 1913 e sostenuta poi in termini più radicali daFrederik Burrhus Skinner (1904-90), questa teoria psicologica ha ispirato molti programmi glottodidattici fino grosso modo agli anni Settanta. Il suoi postulati fondamentali sono che (1) la psicologia studia il comportamento e non la mente; (2) le fonti del comportamento sono esterne, nell'ambiente, e non interne, nella mente. Nelle sue formulazioni più radicali e meno condivisibili viene anche sostenuta la posizione (3) che non esiste (e non solo *non è direttamente studiabile*) altra attività mentale al di fuori dei comportamenti.

[4] Propriamente, per usare l'accurata definizione di Voltolini «Taken in their simplest versions, *externalism* and *internalism* are the conceptions according to which, pending on the broad vs. the narrow identification of an intentional state, the content of such a state can legitimately be conceived only either as relational or as non-relational respectively. For externalists, the representational content of an intentional state depends on a reality lying outside the subject of such a state. For internalists, no external object or event which lies or occurs outside a subject's brain (or at most its body) is relevant for the individuation of the content of an intentional state» [corsivi miei].

In altre parole: che da comportamenti possano inferirsi stati mentali non è affatto controintuitivo; inaccettabile è che *solo* da comportamenti possano inferirsi stati mentali: se per la seconda questione Chomsky aveva certo ragione, il suo errore è semmai stato di fare di ogni behaviourismo un fascio, e la linguistica dei corpora ne ha pagato le penalità.

1.3 La tradizione italiana secondo Sabatini. In Italia, dove il behaviourismo è giunto tardi e non ha mai davvero attecchito, la pregiudiziale generativa attiva nei paesi anglofoni non ha quindi mai potuto giocare un ruolo così rilevante. E, anche se le fondamentali esperienze del padre Busa non hanno goduto della notorietà che meriterebbero, c'è anche dell'altro: la linea "empirica", da linguistica dei corpora *avant la lettre*, che Francesco Sabatini (già presidente della Crusca, e tra i più intelligenti storici della lingua italiana) ha ravvisato nella tradizione lessicografica italiana.

Sabatini ha ripetutamente argomentato che il procedimento *corpus based* (per cui cfr. il paragrafo seguente, dove è contrapposto a quello *corpus driven*) sta alla base della storia linguistica italiana stessa, visto che il *Dizionario* della Crusca, che di quella tradizione rappresenta un momento fondante, è proprio stato costruito su testi (l'idea che la norma si ricavi dall'uso non è di solito associata a posizioni "puristiche" e determina la forma assai peculiare che ha assunto il purismo "cruscante" nostrano). Ma non solo, come dice Sabatini, «il fare preciso ricorso ad un *corpus di testi* [e per la differenza "formale" con i corpora propri della moderna linguistica dei corpora, nella loro accezione più tecnica, cfr. oltre § 2.1] è una costante nell'intera nostra tradizione grammaticografica e lessicografica e, in termini ancora più ampi, nella storia delle dispute linguistiche fin dall'epoca di Dante. Una costante che trova la sua ragion d'essere in una condizione particolare, solitamente considerata penalizzante, della nostra lingua: la sua nascita attraverso l'opera di scrittori e la sua lunga permanenza in vita attraverso l'uso scritto, e quindi grazie al continuo sostegno dato da un canone di autori».

La tradizione italiana, quindi, ha tutte le caratteristiche necessarie per assumere quella funzione centrale nella linguistica dei corpora che è stata finora attribuita a quella anglofona.

1.4 La prospettiva *corpus based* da Fillmore al *Corpus Taurinense.*

In altre sedi ho ripetutamente cercato di trarre le fila di questa situazione, partendo dalla duplice considerazione dell'assenza della pregiudiziale behaviourista, dalla presenza di una tradizione empirica autoctona, cui ho associato l'esistenza di una linea *corpus based* e di un ulteriore elemento continuista (cfr. *infra*).

La linea *corpus based* cui bisogna ricollegarsi è stata lanciata (anche se certo non inventata) in un fondamentale articolo dell'inizio degli anni '90 dal linguista americano Charles J. Fillmore; nato nel '29, appartiene ad una generazione che ha vissuto in prima persona molti degli eventi qui narrati: seguace fin da subito della teoria generativa, cui ha portato notevoli contributi, pure è stato insignito nel 2012 del *Lifetime Achievement Award* da parte dell'influente *Association for Computational Linguistics*: segno della sua ragionata equidistanza da entrambe le pratiche. Saggiamente, infatti, diceva che «I have two major observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way. My conclusion is that the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body»[5]. In altri termini, dai fatti di *parole* raccolti in un corpus si

[5] «La prima osservazione da fare è che non penso ci possa essere nessun corpus, per quanto grande, che possa contenere tutte le informazioni sulle aree della grammatica e del lessico inglese che vorrei esplorare; tutto quello che ho visto è inadeguato. La seconda è che ogni corpus, per quanto piccolo, che ho avuto occasione di esaminare mi ha insegnato cose che non potrei immaginare di scoprire in alcun altro modo. La mia conclusione è che i due tipi di linguista hanno bisogno l'uno dell'altro. O meglio, che i due tipi di linguista, ogni qual volta possibile, dovrebbero coesistere nella stessa persona».

può risalire ai loro correlati stati di *langue* (contro i generativisti più ortodossi), anche se certamente non tutti gli elementi di una *langue* saranno contenuti in un corpus (contro i più accesi antigenerativisti sostenitori della pratica *corpus driven*: il linguista non può e non deve dire altro che non sia desunto da un corpus): è l'uso (testimoniato dai corpora), anzi, che fonda la *langue*, anche se i corpora, essendo per definizione finiti (cfr. la definizione rigorosa data nel § 2.1), ne rappresenteranno solo un sottoinsieme, significativo quanto più il corpus sarà stato costruito in modo accorto (gioco nel quale non può non rientrare la famosa introspezione); ciò, naturalmente, all'insegna della migliore tradizione wittgensteiniana. Anziché, quindi, fare di ogni behaviourismo un fascio, questa è una posizione assolutamente ragionevole; che non abbia attecchito, è probabilmente dovuto all'opposizione radicale, manicheistica ed atavica, tra behavourismo e generativismo esistente nel clima culturale degli *States*, cui abbiamo accennato nel § 1.2.

Trasportata nel diverso clima italiano, questa posizione ha ben diverse *chances* di attecchire. A questo trapianto si è inoltre associata l'idea di tracciare una storia di sostanziale continuità con la tradizione della linguistica filologica otto-novecentesca, ravvisando, così, nella linguistica dei corpora non tanto un elemento di discontinuità e rottura come vorrebbe la tradizione angloamericana (di rivoluzione, in realtà, nella storia della linguistica del secondo Novecento, c'è stata solo quella generativa, come che poi la si voglia valutare), quanto piuttosto, appunto, di continuità con una diversa ma fondamentale tradizione.

Si può, infatti, innovare e contribuire a costruire nuove conoscenze anche lavorando all'interno del solco di una tradizione: posizione che, con paragone extra-epistemologico, era stata resa perfettamente chiara nel campo della storia della musica da Schönberg[6] con il suo *Brahms il progressivo* del '33, il cui intento era, nelle sue parole, «dimostrare

---

[6] Arnold Franz Walther Schönberg (1874 – 1951), compositore viennese, si può considerare il padre della musica moderna: tra i primi sperimentatori dell'atonalità, invento il metodo dodecafonico che è alla base dei serialismo integrale del secondo Novececento. Grande teorico musicale, fu anche un'importante voce del nascente sionismo.

che Brahms[7] – il classicista, l'accademico – fu un grande innovatore nella sfera del linguaggio musicale. Che, in realtà, fu un grande progressivo». Spesso si tende, infatti, a pensare il progredire di una disciplina solo nei termini di "rivoluzioni" e drastici cambi di paradigma di kuhniana memoria; ma in realtà ciò spesso avviene tramite un più lento e meno appariscente accumulo di esperienze, in modo graduale, grazie al lento e "nascosto" lavoro fuori dalle luci della ribalta.

Il ruolo della linguistica dei corpora, almeno in questa accezione, è un po' questo: *innovazione nella tradizione!* E se si dovesse tentare una storiografia linguistica del Novecento, è senz'altro vero che un ruolo di primo piano andrebbe assegnato alla rivoluzione generativista, ma accanto ad essa esistono altre trame (linguistica storica e strutturalismo *in primis*) la cui persistenza è rilevante: non solo hanno diritto ad esistere ma possono ben rivendicare anche la loro importanza; e tra queste la linguistica dei corpora potrebbe essere, appunto, il Brahms della situazione. Il paragone schönberghiano di cui sopra può anche essere spinto più in là: la grossa contrapposizione che segna tutta la metà dell'Ottocento tra wagneriani e brahmsiani, "giovani tedeschi" rivoluzionari ed innovatori e "classicisti" conservatori e tradizionalisti, è un artefatto, una montatura polemica, non rispecchiato dalla realtà dei fatti; e che gli stessi protagonisti di quegli anni la sopportassero come tale, oggi ben sappiamo dalla pubblicazione di molti epistolari, da studi biografici accurati e dalla migliore conoscenza di figure "intermedie" come Joseph Joachim Raff, oltre che dalle argomentazioni strettamente musicali accampate per la prima volta da Schönberg.

Il dialogo con i generativisti meno intransigenti è così riaperto, come dimostrano gli stretti rapporti tra le due recenti imprese gemelle sull'italiano antico del *Corpus Taurinense* e di *ItalAnt*, computazionale l'una e generativa l'altra. Questa minore conflittualità ed apertura al dialogo (in cui probabilmente Lorenzo Renzi, da un campo, e Manuel Barbera, dall'altro, hanno avuto una rilevante parte) è precipua

[7] Johannes Brahms (1833 – 1897), il grande compositore nato ad Amburgo e morto a Vienna, è sempre stato considerato, anche se in parte *malgré lui*, l'ultimo importante esponente dell'ala "classicista" del romanticismo musicale.

caratteristica della situazione italiana, e sarebbe impensabile nelle aree anglofone.

2. I concetti fondamentali.

Così delineato il posto ed il ruolo che spetta alla linguistica dei corpora nella storia della linguistica tutta, e considerato come la sua differenzialità rispetto alla linguistica filologica precedente sia da attribuire più al suo strumentario che alla sua sostanza, è ormai tempo di passare in rassegna i suoi concetti tecnici cardinali, quelli, cioè, che fanno della linguistica dei corpora moderna quello che è.

2.1 La definizione tecnica di corpus.

Ed il primo concetto chiave è naturalmente quello eponimo della disciplina: il corpus, l'oggetto precipuo della linguistica dei corpora. Abbiamo più volte fatto riferimento ad una definizione tecnica e stretta, che è un meditato risultato di un'ampia rassegna condotta nel 2007:

> Raccolta di testi (scritti, orali o multimediali) o parti di essi in numero finito in formato elettronico trattati in modo uniforme (ossia tokenizzati ed addizionati di markup adeguato) così da essere gestibili ed interrogabili informaticamente; se (come spesso) le finalità sono linguistiche (descrizione di lingue naturali o loro varietà), i testi sono perlopiù scelti in modo da essere autentici e rappresentativi.

È questa una definizione "architettonica", basata sugli usi prevalenti che della parola *corpus* la comunità dei linguisti di corpora hanno fatto e fanno, ma legata solo a presupposti formali (il formato elettronico, la tokenizzazione, il markup). Se vogliamo aggiungervi una caratterizzazione contenutistica, sostanziale, dovremmo (recuperando osservazioni che abbiamo in precedenza fatte) aggiungere:

> *Linguisticamente*, inoltre, un corpus è una raccolta di atti di *parole*, e dai fatti di *parole* raccolti in un corpus si può risalire ai loro correlati stati di *langue*, anche se certamente non tutti gli elementi di una *langue* saranno contenuti in un corpus: è l'uso testimoniato dai corpora, anzi, che fonda la *langue*, anche se i corpora, essendo per definizione finiti, ne rappresenteranno solo un sottoinsieme.

Si noti, peraltro, che solo la definizione formale permette di distinguere nettamente tra la prassi della linguistica filologica precedente e quella moderna: la caratterizzazione sostanziale è comune ad entrambe, a riprova di quella continuità di cui sopra dicevamo. La vera differenza tra un "pre-corpus" come i corpora tradizionali quali, ad esempio, il *Corpus juris civilis*, il corpus degli oratori attici, la *Raccolta aragonese*, od il *Codice diplomatico longobardo* ed un BNC (*British National Corpus*) od un PPCME (*Penn-Helsinki Parsed Corpus of Middle English*) risiede praticamente solo nelle suaccennate caratteristiche formali.

Una tale definizione strutturale complessiva, inoltre, come già evidenziato nella rassegna menzionata, non si ritrova in genere nella letteratura internazionale, dove si punta perlopiù a caratteristiche meno formali, come la autenticità e la rappresentatività: ed è questo un ulteriore segno del rigore e dell'originalità della tradizione italiana.

2.2 La definizione legale di corpus. Non è osservazione nuova che la legge rincorra la realtà, spesso restandone assai indietro; e questo divario tra mondo legislativo e mondo reale si è venuto naturalmente acuendo con la robusta accellerazione impressa al mutamento dalle nuove tecnologie. I corpora ne sono un buon esempio, perché le qualifiche legali di cui disponiamo (e che in Italia, almeno, fanno riferimento alle nozioni giuridiche di "banca dati", "opera collettiva" ed "opera derivata") solo molto parzialmente ed imprecisamente si possono rimappare sulla definizione corretta di *corpus* che abbiamo dato poc'anzi: propriamente, infatti, l'unica definizione disponibile nella legislazione italiana è quella, genericissima, di "banca dati" contenuta nel dlgs n. 169 del 1999, art. 2 comma 1:

> Raccolte di opere, dati o altri elementi indipendenti sistematicamente o metodicamente disposti ed individualmente accessibili mediante mezzi elettronici o in altro modo.

Nulla di insolito, si dirà, e potrà sembrare bizzarro che aspetti legali occupino una posizione di rilievo in questa introduzione (e che occupino un intiero capitolo, il terzo, anche nel recente, citato, manuale di riferimento di Tony McEnery ed Andrew Hardie), ma il problema aveva molte ricadute nella ricerca, ed era assai sentito anche dalla comunità internazionale. Una proposta di soluzione giuridica, attivamente cercata da Manuel Barbera e dal gruppo torinese, è arrivata solo cinque anni fa: la proposta, corredata di pratici modelli contrattuali, è basata su *Creative Commons* e precisamente sulle licenze *Share Alike* (o *Condividi allo stesso modo*); si tratta di una soluzione italiana ma facilmente esportabile anche all'estero in quanto fondata su schemi internazionali. Sottrarre i corpora dal limbo giuridico (software od opere a stampa?) in cui si trovavano è equivalso a sdoganarli dall'incubo del copyright, riallineando la linguistica dei corpora al più vasto movimento dell'*open source*, così facilitando la circolazione di risultati e risorse. Programma cui si conforma anche la presente *introduzione*.

Un buon esempio dei guasti portati nella ricerca linguistica da questa incertezza giuridica può essere fornito dalla linguistica testuale: impossibilitati al necessario accesso ai testi completi per ragioni di copyright (a volte malposte: la difettosa acquisizione dei diritti ha infatti in passato portato a cautelative, ma legalmente spesso dubbie, restrizioni dei contesti ottenibili in pubblico, quando non a complete secretazioni dei dati), i testualisti si sono perlopiù defilati dalla linguistica dei corpora, come inadeguata alle loro esigenze. Un'importante ricaduta della "soluzione" suaccennata è stata proprio la consistente (ri)appropriazione della linguistica dei corpora da parte di quella testuale, che finalmente può godere dell'illimitata e piena fruibilità dei contesti fino ai testi intieri; il fenomeno è per ora quasi solo italiano, propagato soprattutto dai gruppi di ricerca di Basilea (svizzeri ma italofoni ed italianisti) e di Torino.

2.3 La finitezza. È questa una condizione indispensabile per almeno due ragioni, una (a) epistemologica ed una (b) pratica.

Quanto ad (a), per garantire la scientificità delle proprie asserzioni, è necessario che le osservazioni fatte possano essere ripetibili; il corpus su cui queste sono condotte deve pertanto essere, oltre che pubblico (e cfr. quanto si diceva nel § 1.3 sull'importanza della questione legale), anche ben definito, stabile e delimitato; cosa difficilmente possibile con corpora non chiusi ed in movimento.

Quanto a (b), il grande vantaggio di essere passati ai moderni corpora informatici dalle schedine cartacee dei linguisti-filologi *d'antan* è soprattutto quello di poter compiere agevolmente (e spesso automaticamente) operazioni statistiche sui dati; la statistica, anzi, è spesso diventata il maggiore marchio di fabbrica della linguistica dei corpora. Non bisogna essere dei grandi statistici per immaginare che qualsiasi operazione statistica implica la stabilità del numero dei dati su cui essa si esercita; minimalmente, una percentuale non può che essere *la percentuale di qualche cosa*.

Pure, all'inizio del millennio vi è stata una proposta, che ha avuto grandissimo seguito, di usare il Web come corpus. Storicamente, che si arrivasse all'esplorazione delle risorse web era inevitabile: l'insufficienza quantitativa delle basi di dati tradizionali per affrontare problematiche linguistiche specifiche sempre più complesse, ed il sempre più rapido "invecchiamento" dei materiali da considerare rispetto al continuo evolversi del linguaggio, in relazione alle nuove tecnologie ed a nuovi mezzi di comunicazione legati alla rete, non potevano che portare, negli ultimi anni, al tentativo di rendere l'intera rete Internet una sorta di mega-corpus da cui estrarre informazioni.

La proposta, seppure utile e prevedibile, si scontra però con il problema della finitezza: il WWW è sempre in movimento, non si può considerare né definito (almeno non nel senso di consentire la ripetibilità degli esperimenti) né finito (nel senso di costituire un insieme numericamente dato, su cui si possano fare operazioni statistiche deterministiche). Infatti, al di là dell'uso diretto del *Web as a Corpus* (come suonava il titolo dell'originario e provocatorio articolo di Adam Kilgariff e Gregory Grefenstette), molto spesso si sono ricavati corpora tratti da materiali web, ma in sé perfettamente chiusi, che fotogra-

fano una data porzione temporale della rete: ad esempio per l’italiano tale è il gigantesco itWaC allestito da Marco Baroni.

2.4 Token (l’elemento minimo di un corpus) e type.

Ritornando alla definizione tenica, essa fa esplicito riferimento ad alcuni concetti irriducibili; alcuni sono ovvi (quello del formato informatico) e non meritano inizialmente particolari esegesi; ma alcuni sono meno ovvi, e sono spesso trascurati nella trattatistica: è questo il caso di token e type.

2.4.1 Token e tokenizzazione. Volando molto raso terra, e tanto per iniziare, per *token* si può intendere l’unità minima in cui è diviso il testo elettronico (che, nel caso più semplice e tipico di un corpus di “testo scritto”, caso che useremo qui come campione, per il computer è solo una lunga stringa di caratteri) di cui è costituito il corpus; la *tokenizzazione*, così, è materialmente la serie di operazioni necessarie per rendere ogni “parola” (od elemento significante del testo, come, in direzione intraverbale, i grafoclitici e, in direzione extraverbale, le multiword, cfr. *infra* § 2.4.3) visibile come token dalla macchina, tipicamente individuandolo con spazi prima e dopo: la tokenizzazione è, in altri termini, il requisito davvero minimo perché un insieme di testi si possa considerare un corpus.

Un esempio, tratto dal CT (*Corpus Taurinense*) di italiano antico, può chiarificare l’operazione:

Brunetto Latini, *Tesoretto*, vv. 113-134.	
versione non tokenizzata a stampa (*testo Contini*, *Poeti del Duecento*)	versione completamente tokenizzata (*testo CT*)
Lo Tesoro conenza. Al tempo che Fiorenza froria, e fece frutto, sì ch'ell'era del tutto la donna di Toscana (ancora che lontana ne fosse l'una parte, rimossa in altra parte, quella d'i ghibellini, per guerra d'i vicini), esso Comune saggio mi fece suo messaggio all'alto re di Spagna, ch'or è re de la Magna e la corona atende, se Dio no·llil contende: ché già sotto la luna non si truova persona che, per gentil legnaggio né per altro barnaggio, tanto degno ne fosse com' esto re Nanfosse.	Lo Tesoro conenza . A ÷l tempo che Fiorenza froria , e fece frutto , sì ch' ell' era de ÷l tutto la donna di Toscana (ancora che lontana ne fosse l' una parte , rimossa in altra parte , quella d' i ghibellini , per guerra d' i vicini) , esso Comune saggio mi fece suo messaggio a ÷ll' alto re di Spagna , ch' or è re de la Magna e la corona atende , se Dio no· lli ÷l contende : ché già sotto la luna non si truova persona che , per gentil legnaggio né per altro barnaggio , tanto degno ne fosse com' esto re Nanfosse .

Tav. 1: la tokenizzazione.

Varie strategie sono state elaborate per automatizzarne il più possibile la procedura, da più sofisticati moduli direttamente inseriti nei tagger (cioè nei software di etichettatura, cfr. oltre 2.5) a semplici applicazioni AWK (un linguaggio di programmazione particolarmente adatto a maneggiare stringhe di testo; molto diffuso è anche il Perl); un programma (o modulo di programma) siffatto prende il nome di *tokenizzatore* (in inglese *tokenizer*).

2.4.2 Token e type: l'orizzonte culturale. Concettualmente, però, le cose non sono così semplici, e navighiamo in acque ben più profonde. La prima definizione risale nientemeno che a Charles

Sanders Peirce, che, nei suoi *Prolegomena to an Apology for Pragmaticism* del 1906, ne dava una definizione illuminante, anche linguisticamente; ecco, integralmente, il celebre passo[8]:

> [536] ... Of the ten divisions of signs which have seemed to me to call for my special study, six turn on the characters of an Interpretant and three on the characters of the Object. Thus the division into Icons, Indices, and Symbols depends upon the different possible relations of a Sign to its Dynamical Object. Only one division is concerned with the nature of the Sign itself, and this I now proceed to state.

---

[8] Riporto anche l'ormai classica traduzione (con minori varianti) di Massimo Bonfantini: «4.536 [...] Delle dieci suddivisioni dei segni che mi sono sembrate degne di uno studio speciale, sei riguardano le caratteristiche di un Interpretante e tre le caratteristiche dell'Oggetto. A esempio la divisione in Icone, Indici, e Simboli dipende dalle diverse possibili relazioni di un Segno con il suo Oggetto Dinamico. Una sola divisione si riferisce alla natura del Segno stesso, e ora mi accingo a definirla.
4.537. Un modo corrente per giudicare della quantità della materia contenuta in un manoscritto o in un libro stampato è contare il numero delle parole, seguendo il metodo messo in uso dal dottor Edward Eggleston. Di solito ci saranno una decina di *il* in una pagina, e naturalmente conteranno per dieci parole. Ma in un altro senso della parola "parola" c'è solamente una parola "il" nella lingua; ed è impossibile che questa parola si manifesti sulla pagina o sia udita in un enunciato orale, per la semplice ragione che essa non è una cosa Singola o un evento Singolo. Non esiste, serve solo a determinare le cose che esistono. Una tale Forma definitamente significante propongo di chiamarla *Type.* Un evento Singolo che accade una volta sola e la cui identita è limitata a quell'unico accadimento o Singolo oggetto o cosa che è in qualche singolo luogo in un istante di tempo dato, un tale evento o cosa che sia significante soltanto in quanto occorre e quando e dove occorre, una cosa come questa o quella parola su una singola riga di una singola pagina di una singola copia di un libro, una tale entità mi azzardo a chiamarla *Token.* Un carattere significante indefinito, come a esempio un tono di voce, non può essere chiamato né Type né Token. Propongo di chiamare un tale Segno *Tone.* Un Type per poter essere usato deve essere reso attuale in un Token, che sarà un segno del Type e perciò dell'oggetto che il Type significa. Propongo di chiamare un tale Token di un Type *Occorrenza* del Type. Così, in una pagina ci potranno essere dieci Occorrenze del Type "il"».

> [537] A common mode of estimating the amount of matter in a MS. or printed book is to count the number of words. There will ordinarily be about twenty *the's* on a page and of course they count as twenty words. In another sense of the word "word", however, there is but one word "the" in the English language; and it is impossible that this word should lie visibly on a page or be heard in any voice for the reason that it is not a Single thing or Single event. It does not exist; it only determines things that do exist. Such a definitely significant Form, I propose to term a *Type*. A Single event which happens once and whose identity is limited to that one happening or a Single object or thing which is in same single place at any one instant of time, such event or thing being significant only as occurring just when and where it does, such as this or that ward on a single line of a single page of a single copy of a book, I will venture to call a *Token*. An indefinite significant character such as a tone of voice can neither be called a Type nor a Token. I propose ta call such a Sign a *Tone*. In order that a Type may be used, it has to be embodied in a Token which shall be a sign of the Type and thereby of the object the Type signifies. I propose to call such a Token of a Type an *Instance* of the Type. Thus there may be twenty Instances of the Type [538] "the" on a page.

Si noti peraltro che già Bonfantini, che è il principale, e benemerito, responsabile della diffusione di Peirce in Italia, nella sua versione manteneva inalterati i termini *Token* e *Type* (mentre traduceva *Instance* con *Occorrenza*); vi sono taluni che in italiano hanno invece preferito "tradurre" ed usare la coppia terminologica "occorrenza *vs* forma", rinunciando ai benefici dell'internazionalismo e della multidisciplinarità, ma soprattutto rischiando di creare quell'illusione che i type siano solo la mera classe dei loro token contro cui già il dettato peirceiano era chiaro, e contro cui aveva ulteriormente e così efficacemente messo in guardia un altro grande filosofo e logico, Willard van Orman Quine: le classi, infatti, devono essere oggetti completamente

astratti, mentre le “classi di token” non lo sarebbero abbastanza per i type, con tutte le aporie che l’uso improrio dell’insieme vuoto notoriamente comporta[9].

Inoltre «It is seldom appreciated that *occurrence* is a third thing: not token, but something between. The word *der* has two occurrences in the sentence *Es ist der Geist der sich den Körper baut*; and I speak now of types, not tokens. Tokens occur in tokens, types in types[10]».

Il mantenimento di tale distinzione, terminologica e concettuale, quale essenziale caratteristica di un corpus consente di ancorare la disciplina non solo alla statistica in generale (dove la percentuale di token e type è uno dei calcoli di base) ma anche alla migliore tradizione semiotica, logica e filosofica, all’insegna dell’internazionalismo e di quella sintesi di elementi matematici e linguistici che è caratteristica precipua della linguistica dei corpora (non a caso si è spesso parlato di “informatica umanistica”).

2.4.3 I paradossi della segmentabilità: grafoclitici vs. multiword. La scansione di un testo in token (determinati convenzionalmente in base a *cosa* si vuole che in un corpus sia poi interrogabile) presuppone che le unità di un testo siano sempre chiaramente segmentabili. Usando, a spanna, il concetto ingenuo di *parola*, possiamo facilmente vedere che ciò non è sempre vero, tanto all’interno (una parola come *della* sarà fatta da due token od uno?) quanto all’esterno (come fare a trattare il *ferro da stiro* come *una* unità lessicale, se sono *tre* token distinti?) della parola.

---

[9] Ad esempio: «The postulate can be put thus: *If* a *and* b *are different strings, then the string consisting of* a *followed by* c *differs from* b *followed by* c. If types were the mere classes of their tokens, this would be false. For, if the strings *a* and *b* have actually been written but are destined never to get written with *c* appended, then the two strings with *c* appended would both be the empty class, if construed as the classes of their tokens, and would thus be identical, contrary to the postulate».

[10] «È raramente riconosciuto che *occorrenza* è una terza cosa: non un token ma qualcosa di intermedio. La parola *der* ha due occorrenze nella frase *Es ist der Geist der sich den Körper baut*; ed io parlo ora di type, non di token. I token occorrono in token, i type in type» (traduzione mia).

Il primo problema è probabilmente il più semplice da domare, perché la sua soluzione dipende da una scelta convenzionale: il linguista deve, ossia, chiedersi se linguisticamente davvero gli serve spezzare l'unità della parola in più token, che poi marcherà in modo da renderli distinguibili dai token "naturali". In italiano è questo di solito il problema dei *grafoclitici*, cioè dei clitici che la tradizione grafica unisce alle parole cui si appoggiano anziché tenerli graficamente distinti (e l'italiano ha entrambi gli usi, *dagli* e *gli da*: come pretendere che i due *gli* siano type dello stesso lemma[11], se uno non è neppure tokenizzato?). Per l'italiano antico si è deciso che era opportuno forzare sempre la divisione (*dagli → da ÷gli* come *gli da*, con *gli* e *÷gli* type del medesimo lemma *gli*; e *degli → de ÷gli* con *de* type del lemma *di* e *gli* del lemma *i*), ma per l'italiano moderno è sufficiente la sola prima divisione, dato che esistono ragioni per mantenere compatte le preposizioni articolate. Computazionalmente, basta studiare ed aggiungere un modulo al tokenizer per trattare anche i grafoclitici nel modo voluto: cosa non facile ma certo non impossibile.

Il secondo problema è senz'altro più difficile, almeno da due punti di vista.
(1) Teoricamente, non è affatto detto che la "multiword"[12] esista come categoria linguistica effettiva (alla stregua di "nome", "verbo", ecc.) o

---

[11] *Grosso modo* per *lemma* si intende l'insieme di tutte le forme flesse (di cui il *paradigma* ne è un segmento significativo) che una parola può assumere, e, metonimicamente, la forma che convenzionalmente è chiamata a rappresentare tale insieme: ad esempio, per i verbi italiani, l'infinito; per il greco ed il latino la prima persona del presente indicativo; per le lingue mordvine la prima persona presente plurale, ecc. Con ulteriore estensione metonimica, si intende talora la voce lessicografica (*articolo*) presente per tale famiglia di forme in un dizionario, o, per riduzione, l'entrata lemmatica sotto cui tale voce è indicizzata. L'idea ingenua, quindi, potrebbe essere che il lemma rappresenti la classe di tutti i type; tale idea ricadrebbe però nelle medesime aporie logiche di quella del type concepito come classe di tutti i suoi token; *linguisticamente*, le conseguenze di ciò sono tuttavia meno pericolose, e la cosa si può eventualmente mantenere, sia pure solo come prima approssimazione. Per la *lemmatizzazione*, poi, cfr. oltre il § 2.6.1.

[12] Come più usualmente la comunità dei linguisti di corpora e computazionali la chiama, dall'inglese *multiword unit*, ma mille altre etichette sono state usate per la stessa realtà, come *locuzione* (*congiuntiva*, *preposizionale*, ecc.), (*unità*) *multilessicale*, *polirematica*, ecc.

sia solo la sommatoria statistica (utile soprattutto in lessicografia ed in svariate attività applicate) di molte realtà linguistiche diverse: entrambe le posizioni sono state sostenute, anche se la seconda è la più indiziata. L'unica cosa sicura è che ci sono diversi tipi di collocazioni (inglese *collocations*; così viene chiamata la disponibilità che le parole hanno ad associarsi – "collocarsi" – tra loro) e che può essere utile in un corpus marcarne almeno qualche tipo.
(2) Computazionalmente, il trattamento da adottare è problematico, e se ne sono date soluzioni molto diverse, che vanno dal ricorso ad una particolare combinazione di fasce di annotazione (cfr. oltre § 2.6.3), a quello ad un apposito chunking (cfr. oltre § 2.6.1), alla decisione di non marcare nulla, ed affidare l'estrazione, quando del caso, a strumenti statistici.

2.5 Il markup ed i metadata. La nozione di markup (in inglese *mark-up*), che segna il difficile confine tra testo e metadata, tra fatti segmentali (→ token) e soprasegmentali (→ markup), tra il corpus in sé stesso e la sua organizzazione, è un altro concetto essenziale, e, nella prospettiva di Peirce sopra riportata, potrebbe essere connesso alla nozione di *tone*, ma è invero più vasto (il *tone* ne sarebbe propriamente un iponimo). L'importanza del markup, tra l'altro, può essere ravvisata anche nella centralità che riveste nella codificazione della TEI (*Text Encoding Initiative*), l'importante consorzio internazionale *non-profit* che sviluppa e mantiene uno standard per la rappresentazione di testi in formato digitale (dove, si noti peraltro, la nozione è stata elaborata proprio da un italiano, l'illuminato storico della logica bolognese Dino Buzzetti).

La nozione "ingenua" di markup come metadata sopra accennata (che in definitiva è poi quella accolta dalla iniziativa TEI) è abbastanza semplice da cogliere, ma ne sono stati più volte fatti notare i limiti semiotici. Dal punto di vista dei corpora, però, è più importante introdurre una distinzione binaria: questa è stata variamente intesa come (a) "markup esterno", cui sono affidati i riferimenti del testo che di esso non fanno costitutivamente parte (autore, titolo, genere, capitoli, paragrafi, pagine, righe ecc.) *vs* "markup interno e filologico", cui sono affidate le informazioni di carattere filologico (integrazioni,

espunzioni, ecc.) e testuale (corsivi, prosa, verso, ecc.); (b) in modo parzialmente sovrapponibile ma diversamente fondato *weakly embedded markup* ‘m. (inserito in modo) sciolto’ o ‘non vincolato’ *vs strongly embedded markup* ‘m. (inserito in modo) vincolato’; (c) più computazionalmente, “posizionale” vs. “strutturale”, di solito associati alla nozione di “attributo”, come avviene nella struttura imposta dal più diffuso software gestore di corpora disponibile, il CWB (*Corpus Work Bench*) col suo CQP (*Corpus Query Processor*), dove si distingue tra *positional attributes* (riferiti ad un token, quindi *strongly embedded,* vincolati) e *structural attributes* (riferiti ad un corpus complessivamente, o ad una sua porzione, quindi *weakly embedded*, non vincolati); il markup contenutisticamente esterno, e formalmente strutturale e non vincolato, è infine spesso riferito *tout court* come “metadata”.

Le cose, come evidente, non sono semplici, ed il confine tra testo e metadata, ineludibile concettualmente e sempre tracciabile nella teoria, nella pratica è spesso confuso, perché deciso convenzionalmente, corpus per corpus, dal costruttore del corpus in base alla combinazione delle esigenze di interrogazione e delle restrizioni imposte dal software di gestione del corpus: che, nel caso del CQP, ad esempio, consente la interrogazione diretta dei soli attributi posizionali e non di quelli strutturali. Anticipando in parte quanto svilupperemo nel § 2.6 sulla codificazione informatica, un esempio (il solito estratto del *Tesoretto* di Brunetto, tratto dal CT, che è codificato col CWB) forse potrà chiarire le idee:

```
<author BrunettoLatini>
<title Tesoretto>
<genr Did>
<chapter 001>
<page 0175>
<type verse>
[...]
<s 1429>
<line 263>
Lo          lo          |art.d|          |60,0,4,6,0,0|     V  Did
Tesoro      tesoro      |n.c|            |20,0,4,6,0,0|     V  Did
conenza     cominciare  |v.m.f.ind.pr|   |111,3,0,6,0,0|    V  Did
.           stop        |punct.fi|       |70,0,0,0,0,0|     V  Did
</s>
<s 1430>
</line>
<line 264>
A           a           |adp.pre|        |56,0,0,0,0,0|     V  Did
÷l          il          |art.d|          |60,0,4,6,0,0|     V  Did
tempo       tempo       |n.c|            |20,0,4,6,0,0|     V  Did
che         che         |pd.rel|         |36,0,4;5,6;7,0,0| V  Did
Fiorenza    firenze     |n.p|            |21,0,5,6,0,0|     V  Did
</line>
[...]
</s 1429>
</type verse>
</page 0175>
</chapter 001>
</genr Did>
</title Tesoretto>
</author BrunettoLatini>
```

Tav. 2: il markup.

Il testo è suddiviso in righe e colonne; nella prima colonna (il cui contenuto non può mai essere nullo) sono contenute le posizioni assegnate ad ogni token, e queste posizioni sono incorniciate dal markup strutturale, che prende la forma di “tag” tra parentesi uncinate, <aperti> e </chiusi>, definendo così delle “regioni” cui il tag si applica. La logica cui obbedisce (propriamente XML, *Extensible Markup Language*, sia pure non rigoroso), quindi, non è molto diversa da quella di una normale pagina HTML, come:

```
<HTML>

<HEAD>
<META name="pippo" content="fuffa">
<TITLE>scempiaggini</title>
<LINK rel="stylesheet" href="disney.css" style type=
"text/css">
<STYLE type="text/css">
<!--
-->
</STYLE>
</HEAD>

<BODY>
<H1>Viva Pippo</SPAN></H1>
<H2>I <I>Dicta memorabilia</I> di Pippo.</H2>
<P>Bla, bla, bla... Ecco, dunque, ma, sì, no, però, già,
beh, magari, vediamo, certo, come no?<BR>
Ecco...<BR>
Ecco...<BR>
</P>
</BODY>
```

Tav. 3: la struttura HTML.

Nelle colonne successive alla prima (per maggiore chiarezza l'esempio è stato semplificato: il CT ne ha in realtà 12), invece, vi sono gli attributi associati ad ogni posizione = token (e non ad una intiera regione di token), cioè il markup più vincolato e *strongly embedded*, nella fattispecie tre fasce di annotazione (il tagging, che per convenzione non si considera effettivamente markup: cfr. il § 2.6 seguente, dove sarà spiegato diffusamente) e due di markup vero e proprio (nella fattispecie l'indicazione se ci troviamo in un verso od in prosa [P/V], e l'indicazione del genere letterario, in questo caso "didattico" [Did]).

2.6 Il tagging. Distinto, come dicevamo, dal markup va considerato anche il tagging, che pure di esso è propriamente una delle tante forme, cioè l'associazione ad ogni token di specifici attributi informativi.

Ci sono molte forme possibili di tagging, che si possono esprimere in diverse fasce (cioè le colonne dell'architettura CQP): linguisti-

camente, quelle più frequenti sono per lemma (detta *lemmatizzazione*), per parte del discorso (detta *POS-tagging*), per categoria sintattica (detta *parsing*), e per valore semantico (*sense tagging* od *annotazione semantica*; è questo un tipo di annotazione che ha avuto recentemente uno sviluppo enorme, soprattutto fuori dai corpora, nel web e nelle "reti semantiche"; per queste ragioni, come già abbiamo detto, non ne faremo qui più che un cenno: ci porterebbe, infatti, troppo lontano e richiederebbe, comunque, un manuale separato).

2.6.1 Lemmatizzazione e parsing. La *lemmatizzazione* è l'operazione, lessicograficamente indispensabile, di ricondurre ogni type al proprio lemma, per cui *canta*, *canteremo* e *canterò* sono marcati tutti come type del lemma *cantare*. L'operazione è concettualmente abbastanza ovvia, ma è informaticamente poco domabile, dato che il lessico è notoriamente la parte meno regolata e razionalizzabile di una lingua: i lemmatizzatori (che spesso sono solo una componente dei più complessi POS-tagger) di solito sono semiautomatici (compiono cioè solo una "lemmatizzazione assistitita"), sennò non potrebbero fare altro che marcare come "sconosciuto" quello che già non trovassero *tel quel* nel proprio dizionario di macchina.

Il *parsing* è un'operazione particolarmente diffusa nei corpora di lingua inglese (i risultati in questo senso raggiunti da Geoffrey Sampson col suo corpus SUSANNE non si stenterà a definirli epocali), dove a fronte di una morfologia particolarmente ridotta, è la sintassi a fornire le indicazioni grammaticali determinanti. Non così in italiano, dove in effetti i corpora parsati sono più l'eccezione della norma (non mancano però anche esperienze in questo senso: cfr. oltre § 3.9). A rigore, tra l'altro, un parsing dovrebbe cogliere solo delle unità realmente linguistiche, dai sintagmi alle frasi, e quindi presupporre che teoricamente esista sempre un nodo "frase" (presupposizione che ad esempio Sampson è ben lungi dal fare); individuare solo realtà statistiche dette *chunk* (a volte coincidenti con quelle linguistiche, ma a volte solo collocazionali) è invece un *chunking*. I software per fare ciò sono detti, rispettivamente, *parser* e *chunker*.

2.6.2 POS-tagging. In italiano (e lingue tipologicamente simili), il caso più tipico è infatti quello della annotazione morfosintattica o *POS-tagging*, su cui, in effetti, per l'italiano ed in Italia, si è molto lavorato. Stante la nostra prospettiva consapevolmente italiano-centrica, è su questa forma di tagging che ci concentreremo.

Stabilire un "tagset" (francese "jeux d'étiquette")[13] per le parti del discorso, prima ancora di pensare alla sua "granularità" (cioè a quanto sia dettagliato ed approfondito) od alla sua efficacia computazionale, implica in primo luogo dare una soluzione, per quanto provvisoria, ad un annoso problema linguistico: *quali* (e possibilmente *cosa*) sono le parti del discorso? Che vengano ancora chiamate all'antica *μέρη τῆς λέξεως* (come nella *Poetica* di Aristotele) o *Partes Orationis* (come in tutta la tradizione occidentale fino a ieri) o *Part of Speech* (come modernamente è d'uso), o più sinteticamente *POS* (come consueto nella linguistica dei corpora), il problema delle parti del discorso è infatti tanto vecchio quanto la molteplicità delle sue *labels* suggerisce; tracciarne una storia complessiva, come peraltro è stato tentato, è qui chiaramente fuori luogo; possiamo però chiederci, da questa plurimillenaria tradizione, quale impostazione sia scaturita che si adatti agli scopi della linguistica dei corpora. In *primo* luogo l'esigenza è che le POS siano intese come classi di parole (classi di lessico, morfologicamente definite in base a proprietà combinatorie; ad es. *Nome* e *Verbo*) e non come tipi di costituente (classi sintattiche, definite in base a proprietà sintattiche; ad es. *Soggetto* e *Predicato*), sennò cessa la possibilità di distinguere tra una fascia di tagging ed una di parsing, distinzione, si è detto, certo utile per lingue come l'italiano, anche se meno per lingue come l'inglese (dove il poco di morfologia necessaria può ben essere sussunta nel parsing). In *secondo* luogo che le POS debbano essere pensate come categorie metalinguistiche (descrittive) e

---

[13] Per inciso: in inglese si dispone di due termini distinti per due concetti distinti, *tag* 'etichetta nel senso sostanziale [ed assoluto: ad es. la categoria *nome*]' vs. *label* 'etichetta nel senso materiale [e contingente; ad esempio *no.*, *n.*, *noun*, *nomen*, *NO*, *N*, ecc.]'; in italiano no: donde la necessità di introdurre il termine *tag* ad affiancare il nativo *etichetta*; in altre parole il *tag* è la categoria, e la *label* od *etichetta* solo il nome di tale categoria.

non realistiche; l'alternativa tra le due impostazioni era già lucidamente delineata nel *Cours* (II.iij) di Saussure:

> Qu'est-ce[14] qu'une *réalité* synchronique? Quels éléments concrets ou abstraits de la langue peut-on appeler ainsi? Soit par example la distinction des parties du discours: sur quoi repose la classification des mots en substantifs, adjectifs, etc.? Se fait-elle au nom d'un principe purement logique, extra-linguistique, appliqué du dehors sur la grammaire comme les degrés de longitude et de latitude sur le globe terrestre? Ou bien correspond-elle à quelque chose qui ait sa place dans le système de la langue et soit conditionné par lui? En un mot, est-ce une realité synchronique? Cette seconde supposition paraît probable, mais on pourrait défendre la première.

L'alternativa, dunque, era quella tra concepire il sistema delle parti del discorso (1) come un sistema logico astratto o (2) piuttosto come una realtà *in re* della struttura del linguaggio oggetto, da cogliere nella sua immanenza; e la linguistica dei corpora deve scegliere risolutamente l'alternativa (1), anche perché allestire un corpus è, da un lato, un'operazione di linguistica applicata e non teorica (un corpus deve poter servire a molti utenti, spesso *non linguisti*, e non solo ai linguisti teorici) e, dall'altro, è un progetto di ingegneria linguistica (il corpus deve essere informaticamente processabile).

Quindi, per *etichettare un corpus* (o meglio, tecnicamente, *POS-taggare*) si deve creare un insieme (*set*) di categorie, ossia un *tagset*, che da una parte possano cogliere alcuni aspetti linguistici significativi, e che dall'altra possano essere facilmente usate da qualsiasi utente,

---

[14] Nella classica traduzione di Tullio De Mauro: «Che cosa è una realtà sincronica? Quali elementi concreti o astratti della lingua possono venire chiamati così? Si prenda ad esempio la distinzione delle parti del discorso: su che poggia la classificazione delle parole in sostantivi, aggettivi ecc.? Si fa in nome di un principio puramente logico, extralinguistico, applicato dall'esterno alla grammatica come i gradi di longitudine e latitudine lo sono sul globo terrestre? Oppure corrisponde a qualche cosa che ha il suo posto nel sistema della lingua ed è da esso condizionata? Insomma, è una realtà sincronica? Questa seconda supposizione parrebbe probabile, ma si potrebbe difendere anche la prima».

non necessariamente un linguista di professione (quindi bando a cose come *complementatori*, *pro-frase*, *elementi-Wh*, *predeterminanti*, giustissime, ma che solo un linguista sa, o perlomeno dovrebbe sapere, cosa significano). Inoltre, tale tagset deve essere applicabile informaticamente in modo il più possibile automatico (e quindi basandosi su informazioni soprattutto segmentali). Questa è la ragione per cui un tagset è, come dicevamo, *assolutamente* metalinguistico, in quanto la sua esistenza si giustifica solo in base alla sua adeguatezza a dei fini (ossia a quello che in logica si chiama *principio di tolleranza*), ma anche *impuramente*, in quanto la sua struttura si giustifica anche in base ad argomenti extraliguistici, applicati, e si può realizzare in gradi diversi, massimo nella architettura generale e minimo nelle singole POS.

Si è pertanto proposto che un tagset efficace debba obbedire ad undici principi; e le considerazioni precedenti giustificano almeno i primi quattro ed il sesto:

1 consensualità e neutralità;
2 adeguatezza descrittiva;
3 standardizzazione;
4 praticità computazionale;
5 tag e *labels* EAGLES-compatibili (corollario di 3);
6 ancoramento morfologico;
7 struttura tipata (*hierarchy-defining features*: HDF);
8 evitamento dei *cross-branchings* con gerarchie separate di MSF (*morphosyntactic features*);
9 contenimento dei tag sotto i 70 (corollario di 4);
10 espansione esplicita di ogni tag gerarchico (corollario di 7);
11 ottimizzazione ed univocità delle *labels* (corollario di 5).

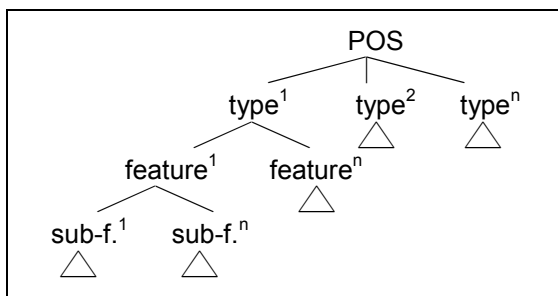
Il principio 5 (che è un'applicazione del terzo) discende da un ulteriore esigenza metalinguistica ed applicata: che i tagset siano progettati in modo da rendere il confronto interlinguistico (cioè la comparazione di corpora di lingue diverse) in massimo grado possibile; EAGLES era un consorzio nato nella metà degli anni '90 proprio per questo scopo, e che aveva prodotto degli standard europei utili ancora adesso.

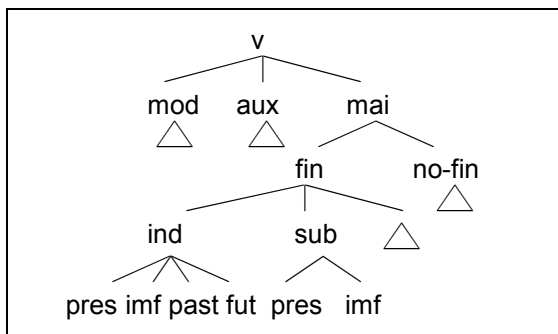
Il principio 7 si rifa invece al concetto di "struttura tipata"; questo tipo di architettura è stata sviluppata in logica all'inizio degli anni '90 e si è presto rivelata particolarmente utile ad organizzare un tagset. Per semplificare, l'idea di base è che i tag siano delle strutture analitiche ad ereditarietà, cioè delle gerarchie di "subtag" in cui ognuno "eredita" le caratteristiche del precedente. Mi spiego con un esempio.

Poniamo che vogliamo etichettare i nomi comuni e propri: potremmo in tal caso ricorrere (1) in una prospettiva tradizionale e "compatta" a due tag le cui etichette potrebbero essere gli usuali *nc* ("nome comune") e *np* ("nome proprio"), o (2) ad unico tag che si identifica con la POS "nome", etichettato *n,* che si suddivide in due *types* (ossia "tipi", donde la qualifica di "tipato" per il sistema[15]), etichettati *com* e *prop*, che potrebbero poi ulteriormente ramificarsi in più *features* e *sub-features*; ogni *com*, in questo caso, "erediterebbe" dal nodo superiore la caratteristica di essere un nome, ecc. Ipotizzando di voler trovare tutte le sequenze di "nome + aggettivo" in un sistema ad etichette gerarchiche (immaginando che *adj* sia l'etichetta del tag "aggettivo") possiamo cercare semplicemente "*n* + *adj*", laddove in un sistema ad etichette compatte dovremmo usare una catena di congiunzioni, tipo "(*np* & *ng*) + *adj*". La maggiore semplicità e duttilità del sistema è evidente, soprattutto quando si pensi a POS molto complesse; l'utilizzo, ossia, di etichette analitiche nella annotazione di un corpus ne permette una descrizione dettagliata e ricerche specifiche, ma l'analiticità risulta dispersiva ed impedisce ricerche generali se non viene sussunta in un sistema di generalizzazioni gerarchiche, fondata sull'ereditarietà. Nell'esempio precedente abbiamo parlato di POS che si suddividono in *types* e quindi in *features* e *sub-features*; in realtà l'approccio definitorio di EAGLES procede piuttosto in senso contrario, *bottom-up*: si parla così di gruppi di *hierarchy-defining features* (HDF), di annotazioni, cioè, che si costruiscono in una gerarchia, e non viceversa; in altri termini, tutte le POS sono la proiezione di un

---

[15] Si badi, peraltro, a non confondere questo tipo con il "type" relato a "token": l'uno, il *type* gerarchico, va considerato termine talmente specifico da doversi o tradurre drasticamente o mantenere inalterato in inglese, e pertanto con plurale in -*s* e stampato in corsivo, laddove l'altro deve essere trattato come prestito non adattato, con plurale invariabile e stampato in tondo.

fascio di *features* gerarchiche (cui ci si riferisce con la sigla HDF); la loro ramificazione (inglese *branching*) più alta è detta *type* e le sotto-ramificazioni (*sub-branchings*) via via più basse sono le *features* POS-specifiche (vedi lo schema seguente); dal punto di vista puramente computazionale, comunque, la questione del verso (*bottom-up* o *top-down*) non è rilevante, in quanto le gerarchie tipate sono percorribili indifferentemente in entrambi le direzioni.

Tav. 4: la gerarchia tipata.

La tavola precedente riproduce lo schema arborescente (oltre al diagramma ad albero sono naturalmente, allo stesso titolo, possibili anche altre modalità di rappresentazione, ad esempio a matrice, ad arco, od a blocchi: si tratta, infatti, comunque di oggetti astratti e distinti dalla loro rappresentazione tipografica) di una “classe di HDF”, che per brevità conveniamo di chiamare semplicemente *HDF*. Per scendere dall’astrazione e proporre invece un esempio concreto, potremmo immaginare per il tag “verbo” una struttura simile, altamente ramificante (uso le *labels* proposte in EAGLES):

Tav. 5: una gerarchia tipata per il “verbo”.

Per il principio 8, bisogna introdurre un’ulteriore specifica: abbiamo visto come trattare le *features* che si proiettano su *una* gerarchia risalendo alla POS lungo un unico percorso, ma non tutte hanno queste caratteristiche. Il genere, ad esempio, non risale ad un’unica sorgente, ma si proietta bensì su più POS o tipi distinti (nome, aggettivo, pronome, participio). In altre parole, gli alberi che descrive si incrociano con molteplici *cross branchings*, vanificando la inequivocità dell’ereditarietà gerarchica. Bisogna pertanto distinguere per queste caratteristiche alcune MSF (*morphosyntactic features*) dalle HDF (*hierarchy-defining features*); nell’architettura EAGLES, che qui si raccomanda, solo le seconde si costruiscono in gerarchia tipata, mentre le prime si applicano liberamente sui tag tipati.

I principi 10 ed 11, infine, sono dei semplici corollari, ma il principio 9, il contenimento numerico dei tag, pur essendo anch’esso propriamente un corollario (nella fattispecie del principio 4, quello della praticità computazionale), non è affatto scontato. I grandi corpora del passato spesso avevano tagset cospicui, e neppure tipati (ad es. il LLC, *London-Lund Corpus of Spoken English*, aveva 197 tag), e così anche alcuni tagset, già tipati, ma recenti (ad es. gli *etiquetaris* usati per catalano, spagnolo ed inglese allo IULA). Questi grandi tagset o erano applicati manualmente o quasi (nella prima fase), o lo sono (semi)automaticamente tramite grammatiche di microregole; l’orientamento prioritario oggi è però in direzione completamente automatica e statistica, ed i tagset devono così essere tagliati in modo da essere appli-

cabili da un tagger stocastico: per fare ciò, si è visto che bisogna limitare il tagset a non più di 70 tag gerarchici.

Questo ci porta a passare dalla parte progettuale e linguistica di un tagset, a quella applicativa ed informatica, che cercherò di contenere all'indispensabile. Teoricamente, un tagset può essere applicato manualmente ad ogni token di un corpus, e questa è più o meno la situazione che si verifica quando si etichetta per la prima volta una lingua (in campo italiano, di questo tipo è stata la grande avventura del CT per il fiorentino tardo-duecentesco); in pratica ciò è molto oneroso, e di fatto inapplicabile per grandi corpora. Praticamente, infatti, è ormai usuale affidarsi ad un apposito software, detto *tagger*, in grado di applicare automaticamente i tag opportuni basandosi su una procedura stocastica, in genere cioè usando i cosiddetti HMM (*Hidden Markov Models*) o "modelli markoviano nascosti", un particolare tipo di modello statistico basato sulla computazione bayesiana di catene di stati, in genere due o tre, particolarmente utile nell'elaborazione informatica del linguaggio orale e scritto. Di tagger stocastici ne sono stati elaborati (e sono tuttora in uso) più di uno; quello più diffuso (sia in ragione della sua efficienza, sia della sua libera distribuzione *open source*) è forse il Tree Tagger, sempre parte del CWB e progettato all'IMS di Stoccarda; ed è alla sua struttura che ci limiteremo.

Questo software opera tanto un'etichettatura per POS quanto una per lemma (lemmatizzazione: *lemma-tagging*), e si compone, essenzialmente, di due moduli base: il programma vero e proprio ed un file di parametri, specifico lingua per lingua. In questo file di parametri sono contenuti (1) un dizionario di macchina (usato per associare i lemmi; nessun calcolo stocastico è in questo caso possibile, le associazioni sono fatte per mera coincidenza di stringhe, cioè *pattern matching*; se il *match*, l'abbinamento, è impossibile la forma viene data come "sconosciuta"); (2) il tagset da usare, limitato alle classi aperte (ché le chiuse dovrebbero essere già ricavabili dal dizionario); (3) un corpus campione già etichettato da usare come controllo dei valori statistici. In pratica per generare un file di parametri bisogna partire da un *training corpus* di circa 250.000 tokens (annotato, di solito manualmente, col tagset desiderato e, almeno teoricamente, "perfetto") ed a partire da questo "ge-

nerare" stocasticamente informazione nuova: nulla si crea dal nulla, e non bisogna pensare alla statistica come ad una pratica stregonesca od alchemica, in grado di trarre l'oro dal piombo (una vecchia massima diffusa tra i linguisti di corpora recitava *garbage in, garbage out*, cioè "quel che ne ottieni è funzione di quel che ci hai messo dentro"). Sul sito dal TreeTagger sono già disponibili due file di parametri italiani (quello di Achim Stein e quello di Marco Baroni), entrambi però con tagset non gerarchici; un terzo, finalmente gerarchico, è da tempo in preparazione a Torino, ma nelle circostanze attuali, i tempi saranno certo ancora lunghi.

2.6.3 Le fasce di annotazione. Un corpus, quindi, può essere semplice (quello che più efficacemente in inglese si chiamerebbe *raw*: essere cioè dotato solo di tokenizzazione e markup) od avere molteplici tipi di tagging (linguistici, filologici, fonetici, ecc.; qui abbiamo sostato soprattutto su quelli più strettamente linguistici, ma il CT che abbiamo usato come corpus *de chevet* ne ha, ad esempio, anche molti filologici), organizzati in altrettante fasce di annotazione.

Ora si potrà, dunque, meglio apprezzare l'esempio (strutturato secondo CQP) di Tav. 2 (tratto da una semplificazione dello schema del CT), in cui, come s'era detto, ogni colonna dopo la prima (obbligatoriamente riempita dai token del testo) è assegnata ad una fascia di annotazione espressa come un *attributo* del token; il tagging è pertanto un particolare tipo di markup, linguisticamente particolarmente importante, individuabile solo in base a ragioni sostanziali e non formali.

Le colonne che ci riguardano in quanto esprimono tipi diversi di tagging in senso proprio sono quelle centrali: la seconda è dedicata alla lemmatizzazione (e presenta quindi la associazione lemmatica del type del dato token); la terza è assegnata al POS-tagging tipato, HDF (ed è espressa da un tag articolato gerarchicamente); e la quarta, infine, presenta un POS-tag non gerarchico e compatto (espresso numericamente) seguito dai valori delle MSF (pure numerici). La compresenza di tag tipati e non è in funzione di una completa flessibilità nell'interrogazione del corpus.

Nell'architettura CQP, quindi, le diverse fasce sono espresse da diverse colonne, ma in altre architetture possono prendere la forma di diversi file, allineati l'uno con l'altro.

2.6.4 Transcategorizzazioni e disambiguazione. Il nodo più problematico che il tagging di una lingua naturale deve affrontare è probabilmente quello delle *transcategorizzazioni*, che, praticamente, è il nome che in linguistica dei corpora (almeno dalla fine degli anni '90) assume il fenomeno che viene ordinariamente chiamato *omografia*: la parola *date* può essere la seconda persona dell'indicativo presente o la seconda persona dell'imperativo del verbo *dare,* così come il plurale femminile dell'aggettivo *dato*, e non solo, potrebbe anche essere il plurale femminile del participio di *dare* così come il plurale del sostantivo *data*; nell'esempio che abbiamo fatto nel § 2.4.3, *gli* può essere ricondotto tanto al lemma *gli* (pronome) od al lemma *il* (articolo); e così via.

Il problema investe tanto la lemmatizzazione (ed allora è stato a volte chiamato *transcategorizzazione esterna*) quanto il POS-tagging (ed allora è stato a volte chiamato *transcategorizzazione interna*), e, se non venisse risolto con un laborioso processo detto *disambiguazione* (ordinariamente con un sistema di regole gestito da uno script, spesso lungo e complesso, vista la sua intrattabilità stocastica), creerebbe notevoli problemi nell'interrogazione di un corpus.

2.7 Codificazione (la rappresentazione del testo). È ora tempo di fermare esplicitamente e tecnicamente qualcosa che avevamo già intuitivamente presentato o dato per scontato. Un testo è (essenzialmente ma non esclusivamente) una successione lineare di stati nel tempo, come già era stato ben ravvisato nel *Cours* di Saussure. Un corpus deve poter *rappresentare computazionalmente* gli elementi del testo in elementi del corpus, e quindi trasformerà sequenzialmente ogni token del testo in una posizione del corpus (una stringa di caratteri separati da uno spazio); gli elementi soprasegmentali o comunque non lineari del testo verranno invece riversati nel markup. Questo può avvenire in modi tecnicamente diversi, ma, limitandoci al sistema del CWB (che teniamo sempre come principale riferimento) avviene asso-

ciando ad ogni posizione, con tutti i suoi attributi, un valore numerico che costituirà l'indice del database su cui (come che siano verbalizzate dall'utente) il software effettuerà materialmente le interrogazioni del corpus. Anche se, ovviamente, una tale rappresentazione puramente lineare è una semplificazione non priva di problemi (non è qui in discussione l'esistenza di elementi paradigmatici o comunque non lineari in un testo), essa può lo stesso utilmente servire come modello di base per una rappresentazione computazionale dei corpora; infatti, è proprio una tale nozione di testo lineare che viene usata in CWB come base della rappresentazione dei testi.

Un esempio di questo sistema per cui il sistema attribuisce numeri alle posizioni dei token di un testo, può essere dato dall'ultimo verso del sonetto XLVII del Cavalcanti (che potremmo assumere anche come monito ad ogni costruttore di corpora), che una volta immesso in un corpus risulterebbe in una successione di 9 token, computazionalmente rappresentati da 9 posizioni, gestite numericamente da una sorta di database:

Fa'	1
ch'	2
om	3
non	4
rida	5
il	6
tuo	7
proponimento	8
!	9

Tav. 6: la rappresentazione numerica.

Naturalmente, l'utente finale nulla vede di tutto ciò, e non ne è forse direttamente interessato, e così neanche il "costruttore ingenuo" di corpora: per quanto riguarda il CQP, esiste un programma ben documentato e con una procedura definita, il cosiddetto *CQP Encoding*, che si occupa di trasformare un materiale testuale in corpus; una con-

sapevolezza di ciò può comunque portare il "costruttore cosciente" a scelte computazionalmente più opportune, cioè informaticamente efficaci.

2.8 Disegno e tipologie di corpora. Ora che abbiamo chiarito quelle cruciali caratteristiche formali e strutturali per cui un corpus è effettivamente tale, possiamo entrare in questioni contenutistiche di disegno e tipologia. Le condizioni generali che avevamo visto poste dalla definizione data nel capitolo 2.1 erano, riassumendo, sostanzialmente: (1) il formato elettronico, (2) la finitezza, (3) la tokenizzazione, (4) la markuppatura.

Ci resta così solo da fare i conti con la clausola condizionale successiva: «se (come spesso) le finalità sono linguistiche [...]». È infatti perentoriamente da osservare che un corpus non deve necessariamente essere *linguistico*: si possono addurre esempi radicali, come il *Canterbury Corpus*, costruito come *benchmark* per valutare gli algoritmi di compressione, o come l'uso "tipo-corpus" che fanno i biologi dei genomi (sequenze "scrivibili" e maneggiabili come una successione testuale di basi); od altri (forse meno radicali) di corpora costruiti con materiali variamente "non testuali" come gli audiovisivi del bellissimo LCCPW (*Lancaster Corpus of Children's Project Writing*).

2.8.1 Autenticità e rappresentatività. Ma restando anche confinati ai tipi di corpora che nessuno si sognerebbe di considerare "non-linguistici", cioè a quelli finalizzati alla descrizione di lingue naturali (o loro varietà), per non avere quella distorsione (*any natural corpus will be skewed...*) un po' pretestuosamente invocata da Chomsky (cfr. § 1.2) sono sempre stati posti due requisiti: (a) la autenticità e (b) la rappresentatività.

Quanto ad (a), si può dire che sia stato il principio guida dell'intera disciplina fin dai suoi albori, e prima (si veda ad es. l'esperienza di Charles Carpenter Fries, soprattutto valorizzata come fa Geoffrey Sampson, cfr. § 1.1); in effetti l'attenzione prestata alla raccolta di dati reali, estratti da una lingua effettivamente prodotta ed usata dai parlanti, è stata subito vista in polemica con ogni esempio studiato a

tavolino da una linguistica “introspettiva”: storicamente, infatti, proprio su tale elemento si snodò il dibattito contro l’intuizionismo generativista e le ormai note obiezioni chomskiane sull’inadeguatezza dei corpora a rappresentare una lingua. L’argomento (al di là degli aspetti “difensivi” storici) non è privo di ragionevolezza, anche se i controesempi non mancano. *In primo luogo*, difficilmente potrebbero passare sotto questa categoria i corpora di comunicazione uomo-macchina (come ad es. ADAM, il corpus di dialoghi annotati per interfacce vocaliche avanzate di Claudia Soria e Vito Pirrelli); la parte spettante alla macchina non è “autentica” nel senso di *naturally occurring language*, ed in questa direzione sarebbe così possibile accettare anche l’idea di non autenticità dei dati per una lingua non prodotta naturalmente dai parlanti ma frutto interamente di dispositivi informatici, come ad esempio in sintesi vocale, generazione automatica di testi o traduzione automatica. *In secondo luogo*, l’altro tipo di eccezione, sempre legata all’elasticità della nozione di “autenticità”, riguarda non tanto la “sinteticità” dei testi quanto la loro “genuinità”: sono ossia testi che variamente si pongono sotto l’insegna della riscrittura, della copia o del plagio; esemplare di questa tipologia è il *METER Corpus*, che si propone proprio di servire come *training corpus* per riconoscimento automatico e misurazione del riuso testuale in àmbito giornalistico. Qualche cautela, nell’invocare l’autenticità, è quindi opportuna.

Quanto a (b), è evidente che, mirando in prospettiva tradizionale all’analisi induttiva di dati linguistici autentici per risalire a conclusioni valide ad un livello più ampio e generalizzato dello studio linguistico, la base empirica debba necessariamente aderire a criteri di rappresentatività, costituisca cioè un campione, un *sample*, della lingua analizzata che ne riproduca idealmente, seppur “in miniatura”, tutte le caratteristiche, pur nell’impossibilità di avere, in ultima analisi, le stesse uguali ed identiche caratteristiche della lingua oggetto di analisi. Questa aporia è ragione delle varie oscillazioni che si sono avute su questo principio, che si sono tradotte di solito nella difficile arte del *bilanciamento* dei corpora, arte tanto indispensabile quanto idiosincratica: un qualsiasi blianciamento, infatti, non può essere tale che in base ad una data finalità, ed essendo le finalità di un progetto di

ricerca virtualmente infinite, lo sono pertanto anche i bilanciamemnti possibili.

2.8.2 Le dimensioni.

Resta da dire di una terza questione, fuori lista ma non meno importante: quella dimensionale. Che i corpora non dovrebbero essere troppo piccoli è requisito ovvio (una massima che figurava in un relativamente faceto decalogo della linguistica dei corpora recitava che «quattro testi che interroghi con la ricerca di Word non sono un Corpus, sono quattro testi»), ma *quanto* piccolo non è effettivamente precisabile.

È un fatto che in tutta la storia della linguistica dei corpora, man mano che la tecnologia lo permetteva, si è avuto un costante aumento (con una robusta impennata nell'era del Web) nella dimensione dei corpora: dal milione (1.014.312 per la precisione) di token del *Brown Corpus*, ai cento e più (110.691.482) del BNC, all'impressionante miliardo e mezzo (1.585.620.279) di itWaC; ciò è naturalmente funzione dell'uso statistico che si fa dei corpora, per cui, ovviamente, più i numeri sono grandi più i calcoli saranno accurati.

Però un dato in controtendenza è la menzionata (cfr. § 2.6.2) dimensione minima necessaria per allenare un tagger stocastico che è di soli 250.000 token. In realtà, inoltre, una volta fatti i conti con questi progetti eccezionali e con la generale tendenza motivata dalla crescita delle disponibilità tecnologiche, ci si scontra con una ben diversa situazione. SUSANNE, ad esempio, la cui importanza non solo per la linguistica dei corpora inglese ma per la linguistica tutta, difficilmente si potrebbe sottostimare, ha "solo" 140.000 parole, (ed anzi SEMiSUSANNE, la versione addizionata di *sense annotation* ne è un'ulteriore frazione); ed il CT, che per l'italiano antico dovrebbe essere quello che SUSANNE fu per l'inglese contemporaneo, ha 259.299 token. Segni tutti che un corpus piccolo, ma ben controllato ed accuratamente annotato, può giocare un ruolo assai importante nello sviluppo della linguistica. Inoltre, su teorizzazione ed uso di corpora di piccole dimensioni soprattutto per la glottodidattica, c'è una ormai notevole tradizione di studi ed una consolidata pratica di cui bisogna pur tenere conto.

L’esistenza, pertanto, di una supposta “soglia minima” sarà da mettere fortemente in dubbio: in altri termini, per dirla con Pietre De Haan[16], «the conclusion seems to be that the suitability of the sample depends on the specific study that is undertaken, and that there is no such thing as the best, or optimum, sample size as such»; la grandezza, cioè, va sempre relazionata alla finalizzazione e le uniche vere soglie minime sono quelle del buon senso e qualla della decenza.

2.8.3 Tipologie di corpora. Le precedenti osservazioni circa la variabilità dei fattori contenutistici, confermano la validità della decisione, in sede definitoria, di attenersi principalmente a caratteristiche formali. Questa variabilità, coniugata all’esistenza, che abbiamo rilevato poc’anzi, anche di corpora non o poco linguistici, porta però alla constatazione che esistono differenti tipologie di corpora, eventualmente con consolidate tradizioni, in base alle discipline coinvolte ed ai materiali trattati.

Pur restando nel seminato, i corpora si potranno così distinguere
(1) in base al medium campionato: corpora testuali (in particolare quelli di lingua scritta), audio (in particolare quelli di lingua parlata) o multimediali;
(2) in base al numero delle lingue coinvolte: corpora monolingui o plurilingui, nel qual ultimo caso potranno essere confrontabili (presentare cioè per ogni lingua testi analoghi, ma non coincidenti) o paralleli (i testi delle varie lingue sono uno traduzione dell’altro, e sono tra loro allineati);
(3) in base al tipo di contenuto campionato: corpora generici bilanciati (tipicamente, i grandi corpora “nazionali”), giornalistici, accademici, giuridici, storici (magari filologici; la categoria è spesso, meno accuratamente, riferita come “corpora diacronici”), dialogici, di vari tipi di CMR (comunicazione mediata dalla rete), di vari tipi di CMC (Comunicazione Mediata dal Computer), di media tradizionali (radiofonici,

---

[16] «La conclusione sembra che la adeguatezza del campione dipende dal tipo di studio che si vuole intraprendere, e che di fatto non esiste in assoluto una dimensione ottimale del campione» (traduzione mia).

televisivi, ecc.), di apprendenti (i cosiddetti *learner corpora*), di interazioni uomo-macchina, di sintesi vocale, ecc.

Come si può ben vedere, i tre criteri base non sono mutualmente esclusivi, anzi si possono benissimo intrecciare: si tratta, in altri termini, di una classificazione teorica, utile a raggiungere una certa chiarezza mentale, ma che poi in pratica, per montare una concreta rassegna di qualche dominio, come quello italiano che schizzeremo nel capitolo 3, va rimontata *ad hoc*.

Viste, peraltro, le finalità didascaliche di questa trattazione, noi ci siamo contentati e contenteremo in questo capitolo di delineare il tipo, per così dire, di base, e cioè quello testuale; ma per la lingua italiana (cfr. oltre § 3) passeremo in rassegna anche le altre tipologie.

2.9 Interrogazione ed espressioni regolari. Visto cosa sono e come sono fatti i corpora, la domanda inevitabile, ora, sarà: come si interrogano? Anche dando per scontato (come abbiamo dato) di non volere affrontare davvero questioni statistiche, pure anche a livello iniziale qualche nozione tecnica sarà necessaria.

2.9.1 Le concordanze. L'operazione più semplice (ed anche una delle prime che sia stata fatta) che si può richiedere trattando di testi è quella di estrarre delle *concordanze*, cioè un elenco alfabetizzato, magari con contesto, di tutte le parole presenti in un testo; per fare ciò, a rigore, non è neppure necessario disporre di un vero corpus, oltre che di un computer: infatti la prima concordanza della storia (quella della *Bibbia Vulgata* curata nel 1262 da Hugo di S. Cher, *recte* Hugo de Sancto Caro, 1200 c. - 1263) era stata fatta a mano, sette secoli e mezzo fa.

Se per Hugo era stata certo un impresa titanica, in epoca informatica questo è un cómpito assolutamente ordinario (effettuabile in ambiente Linux persino da consolle col comando "grep"), automatizzabile anche con semplici script. I software dedicati, detti *concordancer*, non difettano certo e si va da quelli più semplici, come il gratuito SCP (*Simple Concordance Program*), a quelli più complessi come i *WordSmith Tools*, che implicano quasi la realizzazione di un corpus vero e proprio. Naturalmente, con corpora effettivi ed un gestore di corpora come il

CQP si possono avere risultati anche molto più raffinati di una semplice lista in formato KWIC (*KeyWords In Context*)[17].

Quest'operazione, anche se diffusa ed utile, non è certo idiomatica per un corpus, anzi, di solito non vale neppure la pena di costruire un corpus per così poco.

2.9.2 Query ed espressioni regolari. Normalmente ad un corpus si chiede invece di cercare un particolare type (o la cooccorenza di più type), magari con determinati attributi, e magari in un determinato contesto, visualizzandoci (con parametri modificabili) le occorrenze di tutti i suoi token, e magari contandocele. Una simile richiesta, per essere possibile, deve essere formulata in un modo specifico, che usualmente viene chiamato *query*. Il linguaggio in cui vengono espresse le query è quello delle espressioni regolari (o RegExp), anzi una query *è* una RegExp.

Le RegExp sono uno strumento particolarmente potente, e neppure troppo difficile da comprendere ed usare (anche se matematicamente sono fondate in modo rigoroso a partire dall'*algebra degli insiemi regolari* sviluppata dal matematico Stephen Cole Kleene[18] tra gli anni '40 e '50); anzi, alcuni degli operatori sono perfino, in realtà, già inconsapevolmente ben noti ai più, come, ad esempio, l'asterisco <*> che tutti conosciamo per (sia pure un poco impropriamente) 'qualsiasi valore' (<*.*> sta notoriamente per qualsiasi tipo di file, quale sia il suo nome e quale sia la sua estensione); infatti in inglese viene di solito correttamente chiamato non *asterisk*, ma *Kleene star*. Un po' di dimestichezza con le espressioni regolari si raccomanda pertanto a chiunque

[17] In pratica tale formato consiste nella alfabetizzazione di tutti i token del testo e nella loro presentazione allineata con un contesto a destra ed a sinistra di estensione definita. Si tratta di un tipo di indice introdotto da Luhn nel 1960 inizialmente per scopi biblioteconomici, ma presto diventato uno standard per concordanze e corpora.

[18] Aneddoticamente, è curioso che la pronuncia del cognome sia affatto imprevedibile: ['kleɪni:]. Pare che il figlio, Ken Kleene, almeno a quanto riporta (senza dirne peraltro la fonte) la Wikipedia inglese, abbia dichiarato in proposito: «As far as I am aware this pronunciation is incorrect in all known languages. I believe that this novel pronunciation was invented by my father».

sia interessato alla linguistica dei corpora, ed al trattamento automatico di testi in genere: la spesa è poca, e si compra tanto. Ci sono, è vero, versioni lievemente differenti delle RegExp a seconda del linguaggio logico o di programmazione in cui sono usate, ma la questione è di poco conto; qui esemplificheremo parcamente solo la versione implementata in CQP.

Innanzitutto ogni posizione del corpus è rappresentata da una espressione racchiusa tra [quadre]; l'espressione deve dichiarare le coppie attributo-valore desiderate (la posizione 1 è di default chiamata *word*; tutte le altre saranno chiamate con le etichette che sono state dichiarate nel processo di creazione del corpus), ad es.

[word="pinco"],

[POS="nome"].

Tali attributi (che in una semplice query solo per *word* sono anche abbreviabili: *"pinco"*, senza quadre e dichiarazione esplicita, sarebbe possibile, ma *"nome"* no, in quanto troverebbe i *word="nome"* e non le POS volute) sono poi liberamente combinabili all'interno di ogni posizione usando gli appropriati operatori, ad esempio con la congiunzione potremmo avere

[word="pinco" & POS="nome"];

anche le posizioni sono tra loro liberamente combinabili, ad esempio la query (in cui due posizioni sono associate con la concatenazione, espressa dallo spazio, cfr. *infra*)

[word="pinco"] [word="palla"]

mi troverebbe tutte le cooccorenze della parola italiana "pinco" seguita dalla parola "palla". Oltre che ricorrere a designazioni nominali (cioè dichiarazioni dirette, esplicite e singolari) posso esprimere ogni valore con delle variabili ricorsive (che poi altro non sono che i *wildcharacters* o caratteri jolly usati in tutti i motori di ricerca sul web, e cui già siamo avvezzi), di cui le principali sono

.	*dot*	qualsiasi singolo carattere (escluso il *newline* e lo *zero*, 0),
*	*star*	qualsiasi numero di ripetizioni incluso lo *zero*, 0,
+	*plus*	qualsiasi numero di ripetizioni escluso lo *zero*, 0,

normalmente combinate per essere appropriatamente quantificate:

.*	*dot-star*	qualsiasi carattere presente od assente (il *punto* fa match con qualsiasi carattere, e la *stella* permette al punto di essere ripetuto qualsiasi numero di volte, incluso lo zero, cioè anche nessuna volta);
.+	*dot-plus*	qualsiasi carattere per forza presente (il *punto* fa match con qualsiasi carattere, ed il *più* permette al punto di essere ripetuto qualsiasi numero di volte, escluso lo *zero*, cioè almeno una volta).

Ad esempio, la query

[word="pinco"] [word=".*"]

mi troverebbe tutte le combinazioni della parola "pinco" con un'altra parola, quale essa sia (e potrei usare questo risultato per rapportarlo al precedente e calcolare così la disponibilità collocazionale di *pinco* e *palla*).

I principali operatori (anche questi logicamente abbastanza scontati) sono invece

&	*and*	congiunzione (*e*),
|	*or*	disgiunzione (*o*),
=	*value statement*	identità (*uguale*),
!=	*value negation*	negazione di attributo (*non vale "x"*),
!	*not*	negazione (*non*),
<*sp.*>	*concatenation*	concatenazione tra più match od espressioni.

Già limitandosi a queste poche, essenziali, informazioni sarà evidente la grande potenza del sistema.

2.10 Interfaccia di interrogazione. I corpora, siano codificati nel CWB od in altri sofware, possono essere interrogati in locale o via web; in entrambi i casi è necessaria un'interfaccia tra il software che gestisce il corpus e l'utente che lo interroga.

Nella configurazione minima e normale, in locale ed in ambiente Linux o Unix, l'interfaccia è data dalla stessa linea di comando, in cui verrà scritta l'espressione regolare che costituisce la query, il cui risultato verrà stampato sullo schermo, da cui si potrà reindirizzarlo ad un file e/o ulteriormente lavorarlo. *Et c'est tout*. Si tratta della configurazione più semplice, ma anche della più potente, in cui si può sfruttare al completo le molteplici funzionalità del CQP, anche quelle di cui non abbiamo trattato, come i moduli statistici (il CQP può nativamente importare i propri risultati in R, che è l'ambiente *open source* più usato in statistica).

Le interfaccia web pongono intrinsecamente delle limitazioni; pure sono probabilmente la via d'accesso ai corpora oggi più diffusa. In genere i curatori dei siti hanno cercato (a volte con specifiche categorie di utenti in mente) di "semplificare" la ricerca ricorrendo a dei moduli grafici "intuitivi" (un nobile tentativo è ad esempio l'interfaccia dei corpora NUNC preparato da Adriano Allora), ma di solito finendo solo col complicarla ed inutilmente limitarla ancora di più. L'importante, quindi, è che, da sola od accanto ai sistemi grafici, la finestra per immettere i comandi da stringa di tastiera sia sempre presente, come (ad esempio) per il CT, per i NUNC e per il corpus *La Repubblica*; in realtà non serve altro, poi i webmaster possono pure sbizzarrirsi a loro piacimento.

Un caso diverso è quello in cui l'interfaccia web, nonostante le limitazioni intrinseche imposte dalla rete, costituisce invece un valore aggiunto, consentendo anche operazioni non possibili al CQP nativo; qui lo studio delle interfacce, rinunciando ad essere un semplice *maquillage* di talentuosi *web designers*, diventa realmente utile e parte integrante della ricerca. Il migliore esempio di questo tipo è il corpus di apprendenti italiano L2 VALICO (opera prevalentemente di Simona Colombo, diretto da Elisa Corino e Carla Marello, da cui, insieme a Manuel Barbera, fu anche fondato nel 2003); purtroppo, però, esempi analoghi non sono molto diffusi.

3. I corpora disponibili per l'italiano: un panorama.

Con l'idea di concentarmi sulla situazione italiana, volevo anche dare un quadro, rappresentativo sia pure senza pretese di completezza, delle risorse di cui ci può avvalere per questa lingua. I corpora di italiano, soprattutto quelli prodotti nell'ultimo decennio, coprono ormai tutte le principali varietà diamesiche della lingua ed alcune di quelle storiche: si va, per un corno, dallo scritto, al parlato ed alle più diverse forme dei media (italiano degli SMS, dei blog, di Usenet, trasmesso, ecc.), e per l'altro, dalla lingua contemporanea all'italiano del Duecento.

Non tutti i corpora però sono facilmente e gratuitamente accessibili; anzi, la più parte è probabilmente rimasta nel cassetto (cioè nel hardisk) del suo creatore. I limiti (naturalmente applicati con la dovuta elasticità) di questa rassegna saranno pertanto dettati, al di là della effettiva pertinenza dei prodotti censiti ad una definizione stretta di corpus (quella data nel § 2.1), soprattutto dalla loro reale ed effettiva messa a disposizione pubblica (stante anche la centralità della questione legale discussa nel § 2.2), ad esclusione, quindi, di quanto sia proprietario, commerciale, o comunque non accessibile. Questa scelta a favore dell'*open source* e dell'accessibilità pubblica è peraltro coerente con il censimento della linguistica dei corpora italiana promosso dalla SLI nell'ultima delle sue rassegne decennali.

La presente panoramica sarà soprattutto tipologica, sottintendendo la schematizzazione di massima qui suggerita nel § 2.8.3, ma, come là si diceva, rimontandola secondo opportunità.

3.1 Corpora nazionali e bilanciati. Corpora "nazionali", cioè generali, grandi e ben bilanciati, rappresentativi di tutte le varietà "accettate" di una lingua contemporanea, ed in quanto standard liberamente accessibili via Web, sull'onda del BNC (*British National Corpus*), cui ora sta facendo séguito anche il corrispondente ANC (*American National Corpus*), sono ormai disponibili per molte lingue euro-

pee (ad esempio ceco, greco, croato, ungherese, polacco, russo, tedesco, ecc.).

In questo articolato panorama l'italiano sembra essere sorprendentemente assente (in compagnia, sia pure, delle altre due grandi lingue neolatine, lo spagnolo ed il francese): in effetti, è proprio questa la risorsa di cui più si sente la mancanza.

Beninteso, esiste un corpus che pretenderebbe di colmare almeno in parte (limitatamente allo scritto) questa lacuna, il bolognese CORIS (*CORpus di Italiano Scritto*), ma ciò non è propriamente vero: pur essendo, per quanto si possa vedere, eccellentemente costruito (tecnicamente è un prodotto della eccellente mano di Fabio Tamburini, uno dei migliori ingegneri linguistici presenti attualmente sulla scena italiana), il minimo che si possa dire, infatti, è che è scarsamente fruibile, dato che il suo accesso online, ora liberalizzato ma fino a poco tempo fa subordinato ad una complessa e scoraggiante anche se gratuita procedura di registrazione, è limitato a 300 risultati e restituisce solo indici KWIC con appena 30 caratteri di contesto per parte. Purtroppo, quindi, è scarsamente utile (fuorché ai proprietari ed ai loro amici).

3.2 Corpora multilingui. Di corpora multilingui (paralleli e comparabili) che coinvolgano l'italiano, dalla compulsazione della letteratura in materia si ha l'impressione che ne vengano fatti un buon numero, da quelli molto perfezionati come il CEXI (*Corpus of English X Italian*) di Forlì, ad altri più artigianali. Quasi nessuna, però, di queste risorse è diventata pubblica, e rimane pertanto di scarsa o solo potenziale utilità per la comunità degli studiosi.

La principale eccezione è data da un corpus costruito presso l'EURAC (*EURopean ACademy of Bozen/Bolzano*), un centro di indubbia eccellenza e molto importante, ma che di solito è legato a logiche proprietarie. Si tratta del corpus *LexAlp*: costruito e gestito col CWB, mira, secondo recita la homepage, ad un «raffronto contrastivo tra i linguaggi giuridici utilizzati dagli stati dell'arco alpino, con la successiva armonizzazione dei termini principali per la comunicazione sovranazionale». È liberamente consultabile online, e le lingue coperte sono francese, italiano, tedesco e sloveno.

Vi sono, inoltre, altri corpora che, sia pure multilingui, sono stati qui classificati altrove, in ragione del fatto che le loro caratteristiche principali sono altre: (1) confrontabile italiano ed inglese è pure il BoLC, qui considerato nel § 3.3.3; (2) multilingui (per ora italiano, tedesco, francese, spagnolo ed inglese) sono anche i NUNC trattati nel § 3.4.1; (3) italiano ed inglese è il TUT, qui considerato nel § 3.9; (4) italiano, inglese e spagnolo è EPIC, per cui cfr. il 3.8.2; (5) prevalentemente italiano ed inglese sarà infine il televisivo-interpretariale CorIT trattato nel § 3.8.2.

3.3 Corpora di scritto controllato. È la categoria di corpora più tradizionale, e che più facilmente si può utilizzare come surrogato del "corpus nazionale" (cfr. § 3.1) che non c'è.

3.3.1 Giornalistici. Oltre alle tante annate su DVD o CD-ROM che ormai sono disponibili per molte testate, ma che non sono certo dei corpora, anche se naturalmente si possono ben usare (e sono anche state usate utilmente), i corpora giornalistici rappresentano la categoria portante tra i corpora di italiano scritto controllato.

La risorsa principale è il corpus *La Repubblica*, allestito da Marco Baroni a partire da 16 annate dell'omonimo quotidiano. Indicizzato accuratamente col CWB, di cui la maschera di interrogazione ben conserva la duttilità, con i suoi 326.363.463 token è già di dimensioni assai notevoli, (praticamente 3 volte il BNC, che pure già costituiva un vero traguardo internazionale).

Una risorsa più particolare e modesta è il *Corpus Segusinum*, di Manuel Barbera e Cristina Onesti, basato sull'ebdomadario "La Valsusa", di cui è per ora disponibile solo una beta online, che è il primo di una suite di corpora tesi ad esplorare le testate della stampa regionale; in preparazione sono anche la "Gazzetta di Asti" (*Corpus Hastense*) ed "Il Biellese" (*Corpus Eporediense*).

3.3.2 Accademici. La prosa accademica è un genere abbastanza interessante, ma non molto battuto, almeno in pubblico, dalla ricerca.

Praticamente l'unica risorsa per questa categoria è l'*Athenaeum Corpus*, di Manuel Barbera e Luca Valle, un piccolo corpus di prosa accademica prodotta nell'Università di Torino.

3.3.3 Giuridici. Le banche dati di testi legali, ben note ai giuristi, certo non mancano, ma non sono oviamente qui in conto. La situazione dei corpora non è invece ancora soddisfacente, nonostante la linguistica giuridica sia ormai un terreno ben consolidato, e nonostante il reperimento dei testi sia agevole, grazie alla legge 22 aprile 1941, n. 633, art. 5 che stabilisce che «i testi degli atti ufficiali dello stato e delle amministrazioni pubbliche, sia italiane che straniere» non siano coperti dal diritto d'autore.

La grande promessa per il futuro è *Jus Jurium*, di Manuel Barbera e Cristina Onesti, un corpus (o meglio una suite di corpora) appena avviato che vorrebbe documentare il discorso giuridico esistente in Italia all'inizio del nuovo millennio, in tutti i suoi generi, con speciale attenzione agli aspetti testuali e diplomatici.

Un corpus legale comparabile italiano ed inglese è poi il BoLC (*Bononia Legal Corpus*), per cui valgono le stesse considerazioni fatte nel § 3.1 per il suo germano CORIS (tra l'altro, le pretese ragioni di copyright invocate per limitare l'accesso al corpus sono qui chiaramente, appunto, pretese, stante il cit. art. 5 della legge 633 22/4/1941).

Costituito da testi giuridici è pure *LexAlp* di cui abbiamo già parlato nel § 3.2.

3.4 Corpora dei nuovi media. Quantitativamente (soprattutto dopo gli *exploits* di Marco Baroni) si può ben dire che la lingua dei media, e soprattutto quella della rete, faccia la parte del leone nel panorama dei corpora italiani; a ciò avrà certo contribuito tanto la curiosità tecnologica quanto l'ampia disponibilità testuale.

3.4.1 Rete. Il *Web come corpus*, lo avevamo ben visto, è una delle maggiori tendenze della linguistica dei corpora contemporanea, e la linguistica italiana non ha fatto in ciò eccezione: il WWW è stato

esplorato soprattutto da Marco Baroni, e UseNet (cioè la rete che mantiene i cosiddetti *newsgroup*) da Manuel Barbera.

PAISÀ (*Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati*), di Marco Baroni, è costituito da testi raccolti dal web nel settembre/ottobre del 2010; cautelandosi dal punto di vista legale, sono stati accolti solo testi licenziati sotto *Creative Commons Share Alike*: libero pertanto da copyright di sorta, il corpus è tanto scaricabile quanto agevolmente consultabile online. È di dimensione assai ampia (circa 250 milioni di token), è completamente annotato, e trascende ampiamente le finalità glottodidattiche per cui si dichiara nato.

Se già assai cospicua è la dimensione di PAISÀ, quella del gigantesco itWac è addirittura "esagerata": 1.585.620.279 token! Sempre attinto dal Web (limitatamente al dominio *.it*), itWac è anche POS-taggato e lemmatizzato. Il progetto WaCky (che ha prodotto accanto al gigante italiano anche due gemelli inglese e tedesco), in effetti, è mirato proprio alla costruzione di ancora più grandi corpora a partire dal Web, seguendo sì la descritta tendenza, ma evitando la problematica infrazione alla regola della finitezza del corpus (cfr. § 2.3). Baroni ha senz'altro con questa suite di corpora impresso una forte impennata dimensionale (con tutti i suoi benefici effetti statistici) alla linguistica mondiale. Tutti i corpora WaCky sono già liberamente ottenibili, non è tuttavia prevista un'interfaccia web (e, viste le improbabili risorse server necessarie a gestire una simile mole di dati, ciò non stupisce certo).

I NUNC di Manuel Barbera, anch'essi POS-taggati e lemmatizzati, sono basati sui testi delle gerarchie nazionali di Usenet, scaricate dal 2003 ad oggi. Il progetto, di cui sono già stati pubblicati cospicui risultati, ma che è tutt'ora in corso, ha per risultato una innovativa suite di corpora multilingui, anche se l'italiano vi ha avuto sviluppo privilegiato. Una delle caratteristiche più interessanti dei newsgroup è che nascono sempre dal basso in base alla iniziativa degli utenti stessi: la decisione di quali tematiche debbano ricevere una propria bacheca, e di come le bacheche si organizzino all'interno di un dato dominio (nazionale o linguistico) non è decisa dall'alto da una qualche autorità (ministeri, accademie, "specialisti" o *lobbies* di varia natura); l'effetto è che una gerarchia geonazionale di newsgroup si presenta così come

una sorta di "enciclopedia popolare" di una data cultura, un vero ritratto spontaneo della società che l'ha prodotta. Ciò la rende, linguisticamente, di speciale interesse lessicografico, tanto per lo studio dei neologismi, quanto per quello dei lessici specialistici.

3.4.2 Altri media. Anche se per molti di essi (ad esempio per i testi delle segreterie telefoniche) non mancano completamente gli studi, specie di provenienza pragmatica, i corpora sono tutt'ora scarsi.

L'eccezione più rilevante, anche se ancora in corso, sono gli *SMS Monitor Studies* di Adriano Allora, un corpus di SMS al momento di soli 1.394 messaggi, ma in crescita, e già interrogabile.

3.5 Corpora di media tradizionali. A partire dagli anni '80, dopo un fondamentale intervento di Francesco Sabatini, ci si suole di solito riferire a questa famiglia di varietà come *italiano trasmesso*.

Varietà che sono state molto studiate (l'italiano televisivo dispone ormai perfino di un portale dedicato sul Web), ed i cui riflessi nei corpora, nonostante qualcuno non rientrerebbe strettamente in questa rassegna (l'uno perché scomparso, l'altro perché commerciale), si possono considerare complessivamente soddisfacenti.

3.5.1 Televisivi. Il CiT (*Corpus di Italiano Televisivo*) di Stefania Spina fino a non molto tempo fa era consultabile online ma è ora definitivamente scomparso dal Web (il suo dominio risulta in vendita); il che è un peccato, perché, anche se piccolo, era annotato finemente ed in modo accurato.

Per fortuna ve n'è un valido successore, il LIT (*Lessico di Italiano Televisivo*), diretto da Nicoletta Maraschio ed interrogabile online. Raccoglie un campione rappresentativo dell'italiano televisivo del 2006, consistente in 168 ore di parlato tratti dalle reti RAI e Mediaset.

Il Dia-LIT, infine, vorrebbe estendere la campionatura del LIT all'intera storia dell'italiano televisivo, nella sua diacronia dal 1954 ad oggi. In fase di implementamento, una parte ne è già disponibile alla consultazione.

Televisivo, infine sarà anche il CorIT (*Corpus di Interpretazione Televisiva*) trattato nel § 3.8.2.

3.5.2 Radiofonici. In questo caso, purtroppo, la risorsa fondamentale, una e bina, non è libera.

Il LIR (*Lessico di italiano radiofonico*) di Stefania Stefanelli, infatti, non è disponibile online, ma è contenuto in due DVD pubblicati commercialmente dall'Accademia della crusca. Propriamente si tratta di due subcorpora diacronicamente distinti, uno (LIR1) raccolto nel 1995 e l'altro (LIR2) nel 2003. LIR1 consta di circa 64 ore di parlato radiofonico, trascritto e in voce, tratto da nove radio a diffusione nazionale; LIR2 consta invece di 36 ore ed è limitato alla tre reti RAI.

3.6 Corpora storici. Forti di una ricca storia della lingua, vantiamo ormai anche una ricca tradizione di corpora storici, ma purtroppo quasi (cfr. sotto la eccezione del CEOD e quella della *Crusca* online) solo per la fase antica dell'italiano. Naturalmente, non mette qui conto parlare della ben nota LIZ (*Letteratura Italiana Zanichelli*), in quanto commerciale e non in forma di corpus, e neppure delle biblioteche di testi liberi, come Liber Liber (il vecchio Progetto Manuzio). Diverso statuto ha la Biblioteca del CIBIT, da cui non si possono scaricare testi, ma in cui si possono fare semplici ricerche, analogamente a quanto fattibile su Google Libri: non si tratta però di "corpora" in senso pieno, ma solo di utili banche dati testuali.

Innanzitutto va menzionata la banca dati dell'OVI (*Opera del Vocabolario Italiano*), un grandioso e fondamentale database testuale di italiano antico; liberamente consultabile, mantenuto dall'OVI e diretto da Pietro Beltrami, propriamente non rientrebbe (anche se il suo nome ufficiale è *Corpus TLIO*, in quanto base dati per la compilazione del fondamentale *Tesoro della Lingua Italiana delle Origini*) nella stretta definizione data nel § 2.1, ma la sua importanza ed indispensabilità è tale da far passare in second'ordine ogni questione definitoria.

Il CT (*Corpus Taurinense*), poi, è un corpus di italiano antico (nel senso, datogli da Lorenzo Renzi, di fiorentino del secondo Duecento),

ormai giunto alla sua seconda ed ampliata versione (CT+ o neo-CT: disponibile alla medesima homepage). Di modeste dimensioni (attualmente 270.872 token nel CT+) ma accuratamente e riccamente annotato oltre che ampiamente documentato), rappresenta la punta di diamante della sperimentazione di Manuel Barbera (che, con la collaborazione determinante di Marco Tomatis, è il responsabile del progetto) e dovrebbe istituire, nella storia della linguistica dell'italiano antico e nella costruzione di corpora storici, un sicuro standard. In ragione della sua accuratezza, lo abbiamo spesso usato nelle pagine precedenti come fonte di esempi; circa alla funzione che un corpus piccolo ma ben fatto può giocare nella nostra disciplina, abbiamo argomentato nel § 2.8.2.

Un'altra risorsa per l'italiano antico, accessibile ed assai curata, è il *DanteSearch* diretto da Mirko Tavoni a Pisa: comprende tutte le opere di Dante, annotate anche sintatticamente (*Commedia*, *Convivio* e *Rime*) con raffinatezza e dovizia.

La principale eccezione alla "medioevalità" pressoché esclusiva è costituita dall'ottocentesco CEOD (*Corpus Epistolare Ottocentesco Digitale)*, un corpus, coordinato da Massimo Palermo all'Università di Siena, che raccoglie (secondo gli ultimi dati del sito) 1292 lettere, spesso inedite, di 73 scriventi, diversi per provenienza ed estrazione sociale. Interessante anche per le problematiche filologiche spesso affrontate, è completamente accessibile online.

L'altra eccezione è data dalla *Lessicografia della Crusca in Rete*, in cui tutte le cinque edizioni del *Vocabolario della Crusca* sono consultabili e ricercabili online. Come per la banca dati dell'OVI, non si tratta in realtà di un corpus in senso proprio, ma la sua importanza è tale che non se ne può tacere: è infatti eccezionale sia per la sua rilevanza lessicografica in sé, sia per la speciale posizione che la *Crusca* occupa nella tradizione della linguistica dei corpora italiana (cfr. § 1.3).

3.7 Corpora di varietà speciali. Raccolgo sotto questa categoria di comodo i corpora costruiti a partire da tipologie testuali speciali, o comunque meno ordinarie: quelle infantili e quelle dialogiche.

3.7.1 Infantili. Il progetto CHILDES (*CHIld Language Data Exchange System*) è internazionale ed assai importante, pure se parte da interessi più psicologici che linguistici: è stato, infatti, fondato da Brian MacWhinney per studiare il linguaggio infantile, in ogni lingua. Tra le molte lingue in cui si articolano le sue risorse (che, peraltro, a rigore non rientrerebbero strettamente nella nostra definizione di corpus), tutte preparate in CLAN (un programma *free* concepito appositamente per CHILDES) ed agevolmente scaricabili, vi è anche l'italiano.

3.7.2 Dialogici. Il dialogo, anche se è una forma di interazione verbale ormai ben studiata linguisticamente, non si può dire che sia ben rappresentato nei corpora italiani, non essendovi nulla di completamente disponibile.

Del progetto ADAM, di Vito Pirrelli e Claudia Soria, almeno, sono disponibili le specifiche, che sono di prima qualità. Secondo recita la homepage (di cui bisogna necessariamente contentarsi, non potendosi usare il corpus stesso) ADAM sarebbe «un corpus di dialoghi uomo-uomo e uomo-macchina raccolti nel dominio turistico e relativi, rispettivamente, a prenotazioni ed informazioni turistiche e richieste di informazioni sul servizio ferroviario nazionale. Il corpus consiste di 450 dialoghi, ognuno dei quali è rappresentato sotto forma di trascrizione ortografica e di annotazione prosodica, morfosintattica, semantica e pragmatica. Ogni dialogo è inoltre associato ad un file audio che ne registra il segnale».

3.8 Corpora didattici. Quello didattico, soprattutto declinato come corpora di apprendenti (*learner corpora*), è un settore in notevole espansione, un po' come l'apprendologia tutta: le iniziative veramente pubbliche qui riferite sono solo la punta dell'iceberg di una pratica che è assai vasta, anche a condizioni minimali.

Oltre i corpora sotto menzionati bisogna inoltre ricordare che anche altri corpora qui schedati altrove hanno dichiarate finalità didattiche (così PAISÀ: cfr. § 3.4.1) o traduzional-didattiche (così CEXI, cfr. § 3.2).

3.8.1 Di apprendenti. Il progetto più cospicuo è senz'altro il già menzionato corpus di italiano L2/LS VALICO (*Varietà di Apprendimento della Lingua Italiana Corpus Online*), dotato anche di un corpus di controllo L1 VINCA (*Varietà di Italiano di Nativi Corpus Appaiato*). Nati nel 2003 in bmanuel.org e migrati dal 2010 su un dominio indipendente (http://www.valico.org/), sono ora ad esclusiva cura di Carla Marello ed Elisa Corino. Interessano i linguisti applicati ed i glottodidatti perché presentano una grande cura ed abbondanza soprattutto nel trattamento dei metadata sociolinguistici; si segnalano inoltre per quell'attenzione all'interfaccia di cui si è detto nel § 2.10.

Un'altra risorsa di questo genere è LAICO (*Lessico per Apprendere l'Italiano. Corpus di Occorrenze*), coordinato a Siena da Andrea Villarini; il corpus non è al momento interrogabile online, ma lo si può comunque fare scrivendo direttamente all'autore. LAICO raccoglie, per usare le parole della homepage del progetto, 300.516 «occorrenze sulle parole (comprese le polirematiche) presenti nei materiali didattici per insegnare italiano a stranieri. [...] Tutti i testi sono stati archiviati per intero per la loro successiva trattazione lessicometrica con un'accurata indicizzazione che consente di interrogare il corpus in base a vari parametri».

Non liberamente accessibile, ma almeno pubblicato su DVD insieme ad un volume cartaceo, segnalo ancora l'ADIL2 (*Archivio Digitale di Italiano L2*) di Massimo Palermo.

3.8.2 Traduzionali od interpretariali. Tra le varie iniziative attivate, l'unica già disponibile è legata alla SSLMIT di Forlì ed è EPIC (*European Parliament Interpreting Corpus*), un corpus trilingue (italiano, inglese e spagnolo) di testi del Parlamento europeo, allineati e POS-taggati.

Il DIRSI-C (*DIRectionality in Simultaneous Interpreting Corpus*) di Claudio Bertazzoli, di analoga provenienza, non sembra al momento ancora disponibile.

Legata invece alla SSLMIT di Trieste è una risorsa imminente (presto online e consultabile anche al di fuori della Scuola), e che si preannuncia di notevole interesse: il CorIT (*Corpus di Interpretazione Televisiva*), propriamente un corpus multimediale televisivo, che sarà

composto di circa 2.700 *items*, ossia di registrazioni di programmi televisivi italiani in cui sia presente un interprete, ottenute cercando di riunire la maggior parte delle apparizioni di interpreti in TV, attingendo agli archivi della RAI e registrando anche dai canali commerciali italiani; i testi prodotti dagli interpreti sono ovviamente in italiano mentre le lingue di partenza sono diverse con ampia preponderanza dell'inglese.

3.9 Treebank. Il parsing sintattico è stato variamente tentato per l'italiano, ma i risultati sono spesso difficilmente utilizzabili: il VIT (*Venice Italian Treebank*) è disponibile solo commercialmente, e l'ISST (*Italian Syntactic-Semantic Treebank*), che pure sarebbe, con la sua struttura a più fasce, compresa una semantica, forse il più interessante, non lo è neppure a pagamento, secondo una consolidata ma lamentabile prassi dell'ILC (*Istituto di Linguistica Computazionale*) di Pisa.

L'unica risorsa disponibile è quindi il TUT (*Turin University Treebank*), che è dichiaratamente licenziato secondo *Creative Commons Share Alike*, ed è largamente scaricabile. Si tratta di un classico corpus sintatticamente annotato seguendo, analogamente al famoso Treebank di Praga per il ceco, uno schema arborescente a dipendenza, costituito (in base agli ultimi dati presenti sul sito, aggiornato al gennaio 2011) da 2.860 frasi italiane e 200 inglesi, delle cui fonti non è peraltro detto molto, anche se uno ne può indurre che le parti italiane più cospicue siano tratte dal Codice civile e da generici "giornali", in allestimento da anni da parte di un gruppo torinese centrato intorno a Leonardo Lesmo e Cristina Bosco, che hanno pubblicato diffusamente sull'argomento.

Annotato sintatticamente, ma di lingua antica, è poi è l'originale *DanteSearch* che è già stato considerato nella categoria dei corpora storici, § 3.6.

3.10 Corpora di parlato. L'attenzione al parlato ha una lunga tradizione in Italia, rimontando all'impresa lessicografica (peraltro, di lessicografia fondata su corpora nella migliore tradizione britannica)

di Tullio de Mauro del 1993: il corpus del LIP (*Lessico di frequenza dell'Italiano Parlato*), o LIP *tout court*, che ne è derivato è attualmente ancora consultabile sul sito BADIP di Graz (*BAnca Dati dell'Italiano Parlato*).

Il CLIPS (*Corpora e Lessici dell'Italiano Parlato e Scritto*), creato a Napoli da Federico Albano Leoni, è interamente scaricabile previa una semplice registrazione, ed è probabilmente la risorsa oggi di riferimento. È basato su materiali (suddivisi tra radiotelevisivi, dialogici, letti, telefonici ed ortofonici) raccolti in 15 località italiane, oltre che "nazionali", tra il 1999 ed il 2004, presentati in veste sia audio sia testuale.

Se il CLIPS costituisce la più sicura risorsa liberamente disponibile per l'italiano parlato all'inizio del millennio, non bisogna dimenticare che anche al LABLITA (*LABoratorio Linguistico del dipartimento di ITAlianistica*) di Firenze si è lavorato lungamente sul parlato molto e bene. Il C-ORAL ROM, che di queste ricerche è il risultato più cospicuo, non è tuttavia una risorsa libera, anzi è commercializzato a migliaia di euro da ELDA (*Evaluations and Language resources Distribution Agency*); qui la menzioniamo, oltre che per il suo intrinseco valore, perché se non pubblica è almeno "pubblicata" in quanto anche tradizionalmente edita, in veste di libro + DVD.

Un'ultima eccezione, sempre "pubblicata" su CD-ROM in veste editoriale consueta, va fatta, giusta il suo intrinseco interesse, almeno menzionando il corpus di italiano parlato ticinese di PANDOLFI 2007.

4. Bibliografia. Le seguenti indicazioni bibliografiche riguardano tanto i lavori su cui il testo di questo volume è basato, quanto gli ulteriori approfondimenti suggeriti al lettore; non ambiscono minimamente ad essere esaustive ma solo almeno *rappresentative* e, sperabilmente, utili. In questa prospettiva, tra i moltissimi prodotti esistenti, abbiamo sempre privilegiato quegli *open source* o comunque gratuiti. Per snellire il dettato, inoltre, le numerose citazioni e parafrasi da miei precedenti lavori (comunque presenti in bibliografia) non sono tipograficamente segnalate nel testo.

L'assenza della bibliografia dal testo è compensata dal figurare qui in duplice vesta, prima ragionata e poi generale e per esteso.

4.1 Bibliografia ragionata.

4.1 Bibliografia ragionata. L'articolazione seguirà quella dei capitoli del testo: oltre ad illustrare bibliograficamente il testo, saranno qui dati i riferimenti esatti delle citazioni. I riferimenti bibliografici si trovano poi sciolti nella bibliografia generale (§ 4.2).

4.1.0 (Introduzione). La manualistica italiana è ancora praticamente assente; un agile ma efficace profilo è tuttavia BARONI 2010. Quella in lingua inglese invece abbonda: il manuale classico è stato MCENERY - WILSON 2001/1996 ora ottimamente sostituito da MCENERY - HARDIE 2012, destinato a divenire un altro classico; a questi se ne affiancano molti altri, diversi per scopi ed àmbiti, da quello puramente anglistico di MEYER 2002, a quello sociolinguistico di BIBER *et alii* 1998, a quello lessico-terminologico di BOWKER, PEARSON 2002; in altre lingue (tedesca, nella fattispecie) si raccomanda soprattutto l'eccellente LEMNITZER - ZINSMEISTER 2004. I *readers* di prammatica, diversi negli scopi e negli argomenti coperti, ma ugualmente utili e stimolanti, sono MITKOV 2003, SAMPSON - MCCARTHY 2004 e LÜDELING - KYTO 2008-9. Quanto alla statistica i riferimenti essenziali sono MANNING - SCHÜTZE 1999, che si può considerare la vera e propria bibbia della statistica linguistica, ed il più contenuto OAKES 1998, pure assai utile; cfr. anche BARONI - EVERT 2009. Per gli aspetti più computazionali v'è in italiano LENCI - MONTEMAGNI - PIRRELLI 2005. Per i rapporti tra linguistica statistisca e "quantitativa" e tradizionale e "qualitativa" cfr. ancora KLAVANS - RESNIK 1996.

4.1.0.1 (Cos'è in breve la linguistica dei corpora). La citazione di Franco Crevatin è da CREVATIN 2009.

4.1.0.2 (Anglicismi e linguistica dei corpora: un'avvertenza preliminare). BARBERA - MARELLO 2012/03 hanno impostato in un importante convegno dell'Accademia della crusca del 2003 (i cui *Atti* furono però pubblicati solo nel 2012) la questione della terminologia della linguistica dei corpora; Barbera, in particolare, è più volte tornato sulla questione degli anglicismi, cfr. BARBERA 2003 e BARBERA 2007a fino a pervenire alla proposta globale di BARBERA 2009,

§ 1.4, pp. 7-13 (che presenta anche un esempio di lista terminologica dedicata), qui sostanzialmente riassunta ed applicata. Il passo citato delo *Zibaldone* è pp. 3193-6 = ed. PACELLA, pp. 1675-7. Per gli anglicismi in italiano la lettura d'obbligo è ora SABATINI 2011/07, che, oltre a tracciare una efficace storia del "problema", apre delle prospettive "internazionalistiche" cui la soluzione qui prospettata può essere vista come una risposta.

4.1.1. (La linguistica dei corpora nella storia della linguistica: tradizione anglofona *vs* italiana). La citazione è dalla nota 2 di CHOMSKY 1966/2002, a p. 75 della prima edizione e p. 105 della seconda.

4.1.1.1 (La nascita della linguistica dei corpora). La documentazione ancora a stampa (siamo ai primordi...) del *Brown Corpus* è FRANCIS 1964; cfr. anche il manuale online FRANCIS - KUČERA 1979/64. Per Fries, il cui testo base da considerare è FRIES C 1952, probabilmente il suo capolavoro, si veda perlomeno quanto antologizzato in SAMPSON - MCCARTHY 2004; l'inquadramento di riferimento è comunque il sintetico SAMPSON 2004; la bibliografia su Fries è invero vasta, ma si può forse partire dal classico necrologio, MARCKWARTD 1968, uscito su «Language», l'organo della influente *Linguistic Society of America*, per arrivare all'articolo, FRIES P 2010, sulla rivista dell'ICAME, un'altra importante associazione di linguistica dei corpora, questa inglese. Per il padre Busa, la cui importanza fondante è stata riconosciuta già da MARELLO 1996, pp. 167-8, è (metodologicamente e storicamente) determinante BUSA 1951, anche se il risultato finale della sua impresa è BUSA 2005.

4.1.1.2 (Antigenerativismo e tradizione anglofona). Il manuale più classico ed emblematico è quello di MCENERY - WILSON 2001/1996: tutto il primo paragrafo è dedicato alla mossa cui abbiamo accennato; quanto ai testi fondamentali di Geoffrey Sampson si considerino almeno una monografia, SAMPSON 1997, ed una raccolta di saggi, SAMPSON 2001. L'apostolo principale del procedimento *corpus driven* è certo stato il recentemente mancato (2007) John McHardy

Sinclair, di cui cfr. almeno SINCLAIR 1991; per una pacata difesa dell'introspezione, cfr. RENZI 2008/02.

Il fatale intervento di Chomsky è verbalizzato in CHOMSKY 1962/58, e la frase incriminata è a p. 159; un autorevole testimone di quegli anni che ne riporta gli effetti è LEECH 1991, p. 8; la riprova che quasi cinquant'anni dopo Chomsky non abbia cambiato né idea né stile è ANDOR 2004, un'intervista. L'opera di esordio di Chomsky sono le *Syntactic Structures*, CHOMSKY 1957/70, e la recensione antibehaviourista, CHOMSKY 1959/67, è apparsa sul già citato «Language»; oggetto di quella stoccata mortale fu un volume, SKINNER 1953, in cui il più famoso dei behaviouristi, Burrhus Frederik Skinner, aveva condensato decenni di ricerca. In generale, dei rapporti tra generativismo e linguistica dei corpora si è monograficamente occupato BARBERA 2013b; per il behavourismo parafraso in nota BARBERA 2002/10, e l'articolo fondante è WATSON 1913. Per le nozioni di *internismo* ed *esternismo* cfr. VOLTOLINI 1998/2002, da cui è tratta la citazione in nota.

4.1.1.3 (La tradizione italiana secondo Sabatini). Per la definizione della "linea Sabatini" nella tradizione linguistica italiana cfr. SABATINI 2011/06 e 2007, dal quale ultimo lavoro, p. xiij, è tratta la citazione che ho riportato.

4.1.1.4 (La prospettiva *corpus based* da Fillmore al *Corpus Taurinense*). Per la ridefinizione del ruolo della linguistica dei corpora nella storia della linguistica occidentale, cfr. BARBERA 2009, 2011b e 2013. L'importante articolo di Charles J. Fillmore menzionato è FILLMORE 1992, e la citazione è da p. 35; per l'adibizione delle posizioni wittgensteiniane alla linguistica dei corpora, cfr. BARBERA - MARELLO (2008); per l'opposizione tra linguistica *corpus-based* e *corpus-driven* cfr., oltre al già menzionato BARBERA 2013, sia pure con importanti differenze, TOGNINI-BONELLI 2001, pp. 65-100 ed ora MCENERY - HARDIE 2012, capitolo 1.3, pp. 5-6. L'argomento della continuità con la linguistica filologica è stato svolto da Barbera più volte, tra cui più distesamente in BARBERA 2009, p. 23, da cui è tratta gran parte del testo; il riferimento musicale è a SCHÖNBERG 1933/50,

e la sua citazione si trova a p. 60. Sul *Corpus Taurinense* cfr. BARBERA 2009, e per *ItalAnt* cfr. SALVI - RENZI 2010; l'intelligente "apertura" generativista di Lorenzo Renzi è stata RENZI 2008/02 e la risposta di Manuel Barbera, meditata ed a distanza, BARBERA 2013. In generale, per Arnold Schönberg cfr. MANZONI 1975 e per Johannes Brahms SWAFFORD 1997, oltre che, per entrambi, la ricca e pressoché completa discografia esistente.

4.1.2.1 (La definizione tecnica di corpus). Molto di quanto abbiamo detto è basato su BARBERA - CORINO - ONESTI 2007a. In particolare, la definizione formale di *corpus* è tratta da p. 70.

4.1.2.2 (La definizione legale di corpus). Per la situazione legislativa italiana cfr. ZANNI 2007; sull'impatto del problema giuridico nella linguistica dei corpora ha sostato più volte Manuel Barbera, da ultimo 2013*i.s.*, sulla reazione nella comunità internazionale cfr. le discussioni, nei primi anni del Duemila anche molto accese e sconfortate, apparse sulla *mailing list* Corpora, ed ora anche il capitolo 3 (soprattutto pp. 57-60) di MCENERY - HARDIE 2012; per una prima valutazione generale cfr. ALLORA - BARBERA 2007. Il modello di soluzione accennato è fornito in CIURCINA - RICOLFI 2007.

4.1.2.3 (La finitezza). Sui numerosissimi richiami alla finitezza nella letteratura precedente informano BARBERA - CORINO - ONESTI 2007a. Di introduzioni alla statistica ve ne sono mille, ma per la specifica statistica che serve alla nostra disciplina si segnalano i già menzionati OAKES 1998 e soprattutto MANNING - SCHÜTZE 1999. Sulla questione dei *web corpora*, su cui molto si è scritto, si vedano almeno i due poli estremi: da un lato l'articolo che ha lanciato la questione, KILGARRIFF - GREFENSTETTE 2003, pur non essendo certo il primo, si veda almeno VOLK 2002; e dall'altra BARBERA - CORINO - ONESTI 2007a, pp. 44-45, che traggono criticamente le fila della questione. Per gli importanti corpora WaCky di Marco Baroni cfr. infine (oltre alla loro homepage) BARONI *et alii* 2009.

4.1.2.4 (Token e type). Per la concezione più basica di cosa sia un token cfr. GREFENSTETTE - TAPANAINEN 1994; per un esempio (liberamente disponibile) di tokenizer in Perl cfr. il *regexp_tokenizer* di Marco Baroni (vedi homepage). Per AWK si faccia riferimento alla documentazione, ROBBINS 2012, ed alla distribuzione GNU (GAWK) della FSF (*Free Software Foundation*); per Perl cfr. HAMMOND 2003, che è specificamente destinato a linguisti, ed il sito ufficiale del linguaggio in questione (da cui è liberamente scaricabile).

Il lavoro all'origine dei concetti effettivi di token e type è PEIRCE 1933/2006, noto in Italia grazie alla classica antologia di Bonfantini, PEIRCE 1980 (la citazione si trova a p. 230), ora riprodotta anche in PEIRCE 2011 (e la citazione vi si trova a p. 220); l'altro importante passo nella definizione logico-concettuale è in QUINE 1987 (le citazioni sono da p. 218).

Per la nozione di *concetto ingenuo* in linguistica cfr. GRAFFI 1991. Per la nozione di grafoclitico cfr. BARBERA 2009, soprattutto pp. 919-923. Per le due opposte concezioni teoriche delle multiword, cfr. per la linguistica DE MAURO - VOGHERA 1966 e per la statistica BARBERA 2009, pp. 923-925; per un loro possibile trattamento con le fasce di annotazione cfr. BARBERA 2009, pp. 925-948, mentre, in generale, per il trattamento statistico delle collocazioni si veda il sito *collocations.de*.

4.1.2.5 (Il markup ed i metadata). Per il concetto di markup il contributo fondamentale è BUZZETTI 1999, cui in prospettiva corpora si può associare BARBERA - CORINO - ONESTI 2007a, pp. 37-44; per la distinzione di *embedding* cfr. RAYMOND - TOMPA - WOOD 1992. Per la TEI, fanno testo le sue *Guidelines*, giunte ormai alla quinta versione, BURNARD - BAUMAN 2011/08. Per il CWB ed il CQP cfr., oltre al sito, CHRIST - SCHULZE 1996. Per l'XML cf. il sito segnalato oltre (§ 4.2.2).

4.1.2.6 (Il tagging). La storia linguistica del concetto di *parte del discorso*, e la sua definizione per la linguistica dei corpora, è stata delineata in BARBERA 2011a. Per il *Cours* di Saussure si fa naturalmente riferimento all'edizione di De Mauro (il passo citato è da II.iij, p. 133 it. = 152 fr.). Per il principio di tolleranza cfr. CARNAP 1937/34, pp.

51-52 e 1974/63, p. 19. Il riferimento storico per il tagging sono GARSIDE - LEECH - MCENERY 1997 e VAN HALTEREN 1999; si veda anche VOUTILAINEN 2003.

Gli 11 principi dei tagset sono stati dati in BARBERA 2011a, p. 132, sommando BARBERA 2007d e 2007e. La questione della comparabilità interlinguistica e dell'internazionalizzazione, che era stato oggetto dell'iniziativa EAGLES, è stata ripetutamente affrontata da Manuel Barbera, cfr. soprattutto BARBERA 2007e; per gli standard EAGLES si faccia soprattutto riferimento a MONACHINI 1996, oltre agli altri materiali presenti sul sito EAGLES. Per la definizione logica delle gerarchie tipate cfr. CARPENTER 1992 e per l'applicazione all'architettura del tagset cfr. BARBERA 2007d. Per la natura astratta dei tag, e per le grammatiche cosiddette "ad unificazione", che su questo principio si basano, cfr. ALLEGRANZA - MAZZINI 2000. Sulle dimensioni del tagset per l'era precedente il *Penn Treebank* informano MARCUS - SANTORINI - MARCINKIEWICZ 1994, soprattutto p. 274; per i tagset del catalano IULA, si veda il loro sito; per le dimensione adatte ad un tagger stocastico cfr. HEID 1998.

Per l'esempio di un tagger (CLAWS versione 4) diverso dal TreeTagger (ma non gratuito) cfr. GARSIDE - SMITH 1997; per il TreeTagger cfr. invece (oltre al sito) SCHMID 1994; per un mapping dei tagset disponibili per il TreeTagger cfr. BARBERA 2007e; per il tagset di Baroni (sinteticamente fornito nella pagina web sotto riferita), e per la sua prospettiva orientata più all'efficacia computazionale che alla granularità linguistica, cfr. BARONI *et alii* 2004; computazionalmente molto interessanti anche le esperienze di Fabio Tamburini, riportate in BERNARDI *et alii* 2006, di ricavare il tagset direttamente dai dati medesimi da etichettare, in una prospettiva *corpus driven*. Per il modello matematico degli HMM cfr. RABINER (1989) e BLUNSOM 2004, con applicazioni anche al parlato.

Per il concetto di *transcategorizzazione* cfr. BARBERA 2009, soprattutto § 6.5, pp. 82-84; il termine, sia pure in inglese, è stato introdotto nella linguistica italiana da MONACHINI 1996, cfr. in ispecie § 2.1.5, p.11. Per la disambiguazione cfr. TOMATIS 2007, che riporta anche la bibliografia precedente.

4.1.2.7 (Codificazione: la rappresentazione del testo). Per il concetto di rappresentazione informatica del testo e per la struttura che assume in CQP cfr. BARBERA 2011b e soprattutto HEID 2007. Per il CQP Encoding la documentazione ufficiale è EVERT *et alii* 2010a.

4.1.2.8 (Disegno e tipologie di corpora). Le questioni sollevate da questo capitolo sono state sviluppate in BARBERA - CORINO - ONESTI 2007a, pp. 46-7 (natura linguistica), 47-48 (autenticità), 49-51 (rappresentatività) e 53-54 (dimensioni).

Per il concetto di *bilanciamento* è fondante BIBER 1993; utile anche la discussione di TOGNINI-BONELLI 2001, pp. 55-57; per il decalogo della linguistica dei corpora cfr. BARBERA 2007b, la massima citata è la 2.3. Per i vari corpora citati cfr. in genere le rispettive homepage; inoltre per il *Brown Corpus* cfr. FRANCIS - KUČERA 1979/64, per SEMiSUSANNE cfr. POWELL 2006, e per il CT cfr. BARBERA 2009.

Per le dimensioni necessarie ad un *training corpus* cfr. HEID 1998; per l'uso di corpora "piccoli" in glottodidattica cfr. ASTON 1995 e 1997, TRIBBLE 1997, e GHADESSY - HENRY - ROSEBERRY 2002; in generale per le dimensioni dei corpora cfr. DE HAAN 1992, di cui la citazione parafrasata è a p. 3.

4.1.2.9 (Interrogazione ed espressioni regolari). Per fare le concordanze si può andare da un semplice script AWK come *WordLister*, a software dedicati via via più complessi come SCP (*Simple Concordance Program*), *AntConc*, pensato soprattutto con finalità didattiche, cfr. ANTHONY 2004, od i *WordSmith Tools*: per tutti e quattro cfr. le rispettive homepage. Per il formato KWIC cfr. MANNING - SCHÜTZE 1999, § 1.4.5, pp. 31-34; per la sua introduzione, cfr. LUHN 1960.

I manuali di espressioni regolari certo non mancano, ma i più classici sono probabilmente i seguenti: STUBBLEBINE 1993, FRIEDL 2006/1997, e GOOD 2004. Un'articolata guida all'uso delle espressioni regolari del linguaggio di query CQP è stata data da BARBERA 2009, § 21.2, pp. 993-1021, per il CT; BARBERA 2012 ne è un'amplificazione; cfr. naturalmente anche la guida ufficiale, EVERT *et alii* 2010b.

4.1.2.10 (Interfaccia di interrogazione). Il capitolo 3 di HEID 1977, pp. 100-5, passa in rassegna alcune delle intefacce web internazionalmente più diffuse per il CQP. Per R cfr. direttamente il sito; si veda anche ZipfR, un pacchetto statistico specificamente pensato per linguisti computazionali, per cui cfr., oltre al sito, BARONI - EVERT 2006. Per i vari corpora citati si vedano le rispettive homepage; inoltre per i NUNC cfr. BARBERA 2011c.

4.1.3. (Le risorse disponibili per l'italiano). Il presente panorama è basato sull'altra rassegna, pur diversamente atteggiata, effettuata per la SLI (*Società di Linguistica Italiana*): BARBERA 2013*i.s*. Per tutti i corpora menzionati il rinvio principale (e non più ripetuto) è da intendersi alle rispettive homepage.

4.1.3.1 (Corpora nazionali e bilanciati). Per la presentazione del CORIS cfr. ROSSINI FAVRETTI 2000b; per Fabio Tamburini cfr. la sua homepage.

4.1.3.2 (Corpora multilingui). In genere per i corpora paralleli cfr. in italiano GANDIN 2009. Per il CEXI cfr. ZANETTIN 2000 e BERNARDINI 2003. Per *LexAlp* cfr. LYDING *et alii* 2006.

4.1.3.3 (Corpora di scritto controllato). I dati sul corpus *La Repubblica* sono tratti da BARONI *et alii* 2009; per il *Corpus Segusinum* cfr. BARBERA - ONESTI 2010.

Come esempi di banche dati giuridiche, variamente commerciali, si possono guardare *InfoLeges*, *InfoIus* e *Juris Data*. Per due campioni emblematici di linguistica giuridica cfr. MORTARA GARAVELLI 2001 e ROVERE 2005. Per una presentazione del BoLC cfr. ROSSINI FAVRETTI 1998. Per *Jus Jurium* cfr. ONESTI 2010.

4.1.3.4 (Corpora dei nuovi media). Per una tipologia delle comunicazioni mediate dalla rete cfr. ALLORA 2005 e 2009; per una valutazione del loro (in particolare di Usenet) rapporto con la linguistica generale cfr. BARBERA - MARELLO 2008.

Per PAISÀ cfr. BORGHETTI - CASTAGNOLI - BRUNELLO 2011; per il progetto WaCky cfr. BARONI - BERNARDINI 2006 e BARONI *et alii* 2009. Per i NUNC, oltre al citato BARBERA - MARELLO 2008, cfr. soprattutto BARBERA 2011c.

4.1.3.5 (Corpora di media tradizionali). Per la nozione di italiano trasmesso cfr. SABATINI 2011/1982.

Per il portale dell'italiano televisivo, cfr. il sito; per il deceduto CiT (*Corpus di italiano televisivo*) cfr. SPINA 2005/00.

I corpora LIR1/2 sono in STEFANELLI - MARASCHIO 2003.

4.1.3.6 (Corpora storici). Per la LIZ cfr. STOPPELLI - PICCHI 2001; per *Liber Liber*, CIBID e *Google Books* cfr. i rispettivi siti. Per il CT (*Corpus Taurinense*) cfr. BARBERA 2008, e per il CT+ cfr. BARBERA 2012. Per il *DanteSearch* cfr. TAVONI 2011 e per la *Crusca online* cfr. BIFFI 2012. Per il CEOD cfr. infine ANTONELLI - CHIUMMO - PALERMO 2004.

4.1.3.7 (Corpora di varietà speciali). Per il progetto CHILDES, oltre al sito, cfr. MACWHINNEY 2000; per il CLAN cfr. il sito, con ampia documentazione.

Per la linguistica del dialogo un buon riferimento può essere BAZZANELLA 2002.

4.1.3.8 (Corpora didattici). In generale, in lingua italiana, cfr. ANDORNO - RASTELLI 2009. Per una storica apologia dei corpora didattici piccoli e fai-da-te (pratica almeno in parte responsabile del pullulare di iniziative personali e private nel settore) cfr. TRIBBLE 1997. Per VALICO cfr. CORINO, MARELLO 2009abc ed ALLORA - COLOMBO - MARELLO 2011; per LAICO cfr. VILLARINI 2008 e 2011. ADIL2, infine, è pubblicato in PALERMO 2009.

Per EPIC cfr. BERTAZZOLI 2010. Per il CorIT cfr. FALBO 2012 e STRANIERO 2007.

4.1.3.9 (Treebank). Per l'ISST cfr. MONTEMAGNI *et alii* 2003; tra le molte pubblicazioni sul TUT cfr. almeno LESMO - LOMBARDO - BOSCO 2002.

4.1.3.10 (Corpora di parlato). Come testo di riferimemento per la tecnica acustica in linguistica computazionale cfr. in genere JURAFSKY - MARTIN 2000. Per il C-ORAL ROM cfr. CRESTI - MONEGLIA 2005.

4.2 Bibliografia generale. Vi si trovano sciolti, e presentati in un unico compatto ordine alfabetico, tutti i riferimenti abbreviati delle parti precedenti. L'articolazione prevedue due sezioni, una dedicata ai riferimenti bibliografici veri e propri, ed una seconda a quelli web (homepage dei corpora menzionati e siti altrimenti di interesse).

4.2.1 Riferimenti bibliografici.

AA. VV.

1994 *Proceedings of the 3rd International Conference on Computational Lexicography (COMPLEX '94)*, Budapest, Research Institute for Linguistics - Hungarian Academy of Sciences, 1994.

1999 *Il ruolo del modello nella scienza e nel sapere. Roma, 27-28 ottobre 1998*, Roma, Accademia Nazionale dei Lincei, 1999 "Contributi del Centro Linceo Interdisciplinare 'Beniamino Segre'" 100.

2002 *Proceedings of the International Conference on Natural Language Processing (ICON 2002)*, Mumbay (India), s.e., 2002.

2004 *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, ELDA, 2004.

2006a *MLRI '06. Multilingual Language Resources and Interoperability. Proceedings of the Workshop, 23 July 2006 Sydney, Australia*, Burwood (AU) - Stroudsburg (USA), BPA Digital - The Association for Computational Linguistics, 2006 "COLING - ACL 2006".

2006b *Proceedings of the 5th International Conference on Language Resources and Evaluation - LREC 2006*, Genova, LREC, 2006.

ABEILLÉ
2003 *Building and using Parsed Corpora*, edited by Anne Abeillé, Dordrecht, Kluwer, 2003 “Language and Speech series”.

AIJIMER - ALTENBERG
1991 *English Corpus Linguistics. Studies in Honor of Jan Svartvik*, edited by Karin Aijimer and Bengt Altenberg, London - New York, Longman, 1991.

ALLEGRANZA -MAZZINI
2000 Valerio Allegranza - Giampaolo Mazzini, *Linguistica generativa e grammatiche a unificazione*, Torino, Paravia, 2000 “Scriptorium. Sapere linguistico e pratica dell’italiano”.

ALLORA
2005 Adriano Allora, *A Tentative Typology of Net Mediated Communication*, comunicazione presentata alla *Corpus Linguistics 2005 Conference, Birmingham July 14-17 2005*, disponibile online alla pagina http://www.corpus.bham.ac.uk/PCLC/.
2009 Adriano Allora, *Variazione diamesica generale nelle Comunicazioni Mediate dalla Rete*, in «Rassegna Italiana di Linguistica Applicata» III (2009) 147-170.

ALLORA - BARBERA
2007 Adriano Allora - Manuel Barbera, *Il problema legale dei corpora. Prime approssimazioni*, in BARBERA - CORINO - ONESTI 2007a, ¶ 5 pp. 109-118.

ALLORA - COLOMBO - MARELLO
2011 Adriano Allora - Simona Colombo - Carla Marello, *I corpora VALICO e VINCA: stranieri e italiani alle prese con le stesse attività scritte*, in MARASCHIO - DE MARTINO - STANCHINA 2011, pp. 49-61.

ANDOR
2004 JÓZSEF ANDOR, *The Master and his Performance: An Interview with Noam Chomsky*, in «Intercultural Pragmatics» I (2004)[1] 93–111.

ANDORNO - RASTELLI
2009 *Corpora di Italiano L2: Tecnologie, metodi, spunti teorici*, a cura di Cecilia Andorno e Stefano Rastelli, Perugia, Guerra Edizioni, 2009.

ANTHONY
2005 LAURENCE ANTHONY, *AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit*, in ANTHONY *et alii* 2005, pp. 7-13, disponibile anche online: http://www.antlab.sci.waseda.ac.jp/research/iwlel_2004_anthony_antconc.pdf

ANTHONY *et alii*
2005 *Proceedings of IWLeL 2004: an Interactive Workshop on Language E-learning, Tokio, Waseda University December 10th, 2004*, edited by Laurence Anthony, Shinichi Fujita; Yasunari Harada; Waseda Daigak, [Tokyo], ,Waseda University, 2005.

ANTONELLI - CHIUMMO - PALERMO
2004 *La cultura epistolare nell'Ottocento. Sondaggi sulle lettere del CEOD*, a cura di Giuseppe Antonelli, Carla Chiummo e Massimo Palermo, Roma, Bulzoni, 2004.

ANTONINI - STEFANELLI
2011 *Per Giovanni Nencioni. Convegno Internazionale di Studi. Pisa - Firenze, 4-5 Maggio 2009*, a cura di Anna Antonini e Stefania Stefanelli, Firenze, Le Lettere, 2011.

ARMSTRONG
1994 Susan Armstrong, *Using Large Corpora*, Cambridge (Mass.) - London (En.), The MIT Press, 1994 "A Bradford Book", "ACL-MIT Press Series in Computational Linguistics" = «Computational Linguistics» XIX (1993)[1-2].

ASTON

1995 Guy Aston, *Corpora in Language Pedagogy: Matching Theory and Practice*, in COOK - SEIDLHOFER 1995, pp. 257-270.

1997 Guy Aston, *Small and Large Corpora in Language Learning*, in LEWANDOWSKA - TOMASZCZYK - MELIA 1997, pp. 51-62; disponibile online alla pagina http://www.sslmit.unibo.it/~guy/wudj1.htm.

BARBERA

2002/10 Manuel Barbera, *Introduzione alla linguistica generale. Corso online*, 29-12-2002_1, 3-1-2004_2, 25-12-2005_3, 1-12-2010_4. Homepage: http://www.bmanuel.org/courses/corling_idx.html.

2003 Manuel Barbera, *Review* to Manfred Görlach, *A Dictionary of European Anglicisms. A Usage Dictionary of Anglicisms in Sixteen European Languages*, in «International Journal of Lexicography» XVI (2003)^2, pp. 208-216.

2007a Manuel Barbera, *La resa dei forestierismi in italiano. Breve nota ortografica*, in BARBERA - CORINO - ONESTI 2007a, ¶ iiij pp. xv-xvj.

2007b *Il decalogo della Corpus linguistics. (Tanto* Esodo *20,2-17 e* Deut. *5,6-21 erano diversi)*, in BARBERA - CORINO - ONESTI 2007a, ¶ 2 pp. 21-23.

2007c Manuel Barbera, *Per la storia di un gruppo di ricerca. Tra bmanuel.org e corpora.unito.it*, in BARBERA - CORINO - ONESTI 2007a, ¶ 1 pp. 3-20

2007d Manuel Barbera, *Un tagset per il Corpus Taurinense. Italiano antico e linguistica dei corpora*, in BARBERA - CORINO - ONESTI 2007a, ¶ 8 pp. 135-168.

2007e Manuel Barbera, *Mapping dei tagset in bmanuel.org / corpora.unito.it. Tra* guidelines *e prolegomeni*, in BARBERA, CORINO, ONESTI (2007a), ¶ 23 pp. 373-388

2009 Manuel Barbera, *Schema e storia del "Corpus Taurinense". Linguistica dei corpora dell'italiano antico*, Alessandria, Edizioni dell'Orso, 2009.

2011a Manuel Barbera, *“Partes Orationis”, “Parts of Speech”, “Tagset” e dintorni. Un prospetto storico-linguistico*, in BORGHI - RIZZA 2011, tomo I, pp. 113-145. Rielaborazione di una lezione inedita, *Parti del discorso ed annotazione di corpora elettronici*, tenuta a Basilea il 9 maggio 2008 presso l’Instituto di Italianistica dell’Universität Basel.

2011b Manuel Barbera, *Intorno a “Schema e storia del Corpus Taurinense”*, comunicazione al *III Incontro di filologia digitale, Verona, 3-5 marzo 2010*, ora in COTTICELLI KURRAS 2011, pp. 27-48.

2011c Manuel Barbera, *Une introduction au NUNC: histoire de la création d’un corpus*, in «Verbum» XXXIII (2011)[1-2], 9-36 = FERRARI - LALA 2011. Una versione italiana è in BARBERA 2013c, pp. 97-114.

2012 Manuel Barbera, *Il neo-Corpus Taurinense e l’arte della query*, comunicazione al *Seminario: sintassi dell’italiano antico e sintassi di Dante. Pisa 14-15 ottobre 2011*, ora in TAVONI 2012, pp. 61-79

2013b Manuel Barbera, *Per una soluzione teorica e storica dei rapporti tra grammatica generativa e linguistica dei corpora*, relazione *7es Journées suisses de Linguistique. L’empirie en linguistique: variété et complexité des approches. Lugano, Università della Svizzera italiana, 13-14 settembre 2012*, poi revisionato in BARBERA 2013c, pp. 27-45; la versione vecchia è online nel “Schweizerische Sprachwissenschaftliche Gesellschaft / Société Suisse de Linguistique (SSG/SSL) – Archive”: http://www.sagw.ch/fr/ssg/taetigkeiten/7e-Giornate-svizzere-della-Linguistica.html.

2013c Manuel Barbera, *Molti occhi sono meglio di uno: saggi di linguistica generale 2008-12*, [Milano], Qu.A.S.A.R.

2013 *i.s.* Manuel Barbera, *Linguistica dei corpora*, in IANNACCARO *i.s.*

BARBERA - CORINO - ONESTI

2007a *Corpora e linguistica in rete*, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007 “L’officina della lingua. Strumenti” 1.

2007b Manuel Barbera - Elisa Corino - Cristina Onesti, *Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup*, in BARBERA - CORINO - ONEST 2007a, ¶ 3 pp. 25-88.

BARBERA - MARELLO
2008 Manuel Barbera - Carla Marello, *Tra scritto-parlato*, Umgangssprache *e comunicazione in rete: i corpora NUNC*, in «Studi di Grammatica Italiana» XXVII (2008) 157-185 = ANTONINI - STEFANELLI 2011.
2012/03 Manuel Barbera - Carla Marello, *Corpo a corpo con l'inglese della corpus linguistics, anzi, della linguistica dei corpora*, in NESI - DE MARTINO 2012, pp. 357-370.

BARBERA - ONESTI
2010 Manuel Barbera - Cristina Onesti, *Dalla Valsusa in avanti: i corpora di stampa periodica locale*, in «Rivista Internazionale di Tecnica della Traduzione | International Journal of Translation» XII (2010), Special issue *Traduzioni nella stampa periodica*, a cura di Stefano Ondelli, pp. 103-116.

BARNI - TRONCARELLI - BAGNA
2008 *Lessico e apprendimenti. Il ruolo del lessico nella linguistica educativa*, a cura di Monica Barni, Donatella Troncarelli e Carla Bagna, Milano, Franco Angeli, 2008.

BARONI
2010 Marco Baroni, *Corpora di lingua italiana*, in SIMONE 2010-11, *sub vocem*, vol. I, pp. 300b-303a.

BARONI - BERNARDINI
2006 *WaCky! Working Papers on the Web as Corpus*, edited by Marco Baroni and Silvia Bernardini, Bologna, GEDIT edizioni, 2006, disponibile online alla pagina http://wacky book.sslmit.unibo.it/.

BARONI *et alii*
2004 Marco Baroni - Silvia Bernardini - Federica Comastri - Lorenzo Piccioni - AlessandraVolpi - Guy Aston - Marco Mazzoleni, *Introducing the* La Repubblica *Corpus: A Large, An-*

notated, TEI(XML)-Compliant Corpus of Newspaper Italian, in AA. VV. 2004, pp. 1771-1774, online a http://www.form.unitn.it/~baroni/publications/lrec2004/rep_lrec_2004.pdf.

2009 Marco Baroni - Silvia Bernardini - Adriano Ferraresi, Eros Zanchetta, *The WaCky wide web: a collection of very large linguistically processed web-crawled corpora*, in «Journal of Language Resources & Evaluation» XLII (2009) 209–226, online a http://wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wacky_2008.pdf.

BARONI - EVERT

2006 Marco Baroni - Stefan Evert, *The ZipfR Library: Words and Other Rare Events in R*, comunicazione al convegno *useR! 2006, Vienna, 15 June 2006*, handout disponibile online a http://www.r-project.org/useR-2006/Slides/Evert+Baroni.pdf

2009 Marco Baroni - Stefan Evert, *Statistical Methods for Corpus Exploitation*, in LÜDELING - KYTO 2008-9, Volume 2, pp. 777-802.

BAZZANELLA

2002 *Sul dialogo. Contesti e forme di interazione verbal*e, a cura di Carla Bazzanella, Milano, Guerini e associati, 2002.

BENDAZZOLI

2010 Claudio Bendazzoli, *Corpora e interpretazione simultanea*, Bologna, Asterisco, 2010.

BENINCÀ *et alii*

1996 *Italiano e dialetti nel tempo. Saggi di grammatica per Giulio C. Lepschy*, a cura di Paola Benincà, Guglielmo Cinque, Tullio de Mauro e Nigel Vincent, Roma, Bulzoni Editore, 1996.

BERNARDI *et alii*

2006 Raffaella Bernardi - Andrea Bolognesi - Corrado Seidenari - Fabio Tamburini, *POS Tagset Design for Italian*, in AA. VV. 2006b, pp. 1396-1401.

BERNARDINI
2003 Silvia Bernardini, *Designing a Corpus for Translation and Language Teaching: The CEXI Experience*, in «TESOL Quarterly» XXXVII (2003)[3] 528-537.

BIBER
1993 Douglas Biber, *Representativeness in Corpus Design*, in «Literary and Linguistic Computing» VIII (1993)[4] 243-57.

BIBER *et alii*
1998 Douglas Biber - Susan Conrad - Rand Reppeni - Jean Aitchison, *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge, Cambridge University Press, 1998 "Cambridge Approaches to Linguistics".

BIBER - FINEGAN
1991 Douglas Biber - Edward Finegan, *On the Exploitation of Computerized Corpora in Variation Studies*, in AIJMER - ALTENBERG 1991, pp. 204-220.

BIFFI
2012 Marco Biffi, *La Crusca si riscatta nel digitale*, in «La Crusca per voi» XLV (2012)[ottobre] 18-19.

BLUMSON
2004 Phil Blumson, *Hidden Markov Models*, lecture notes, 2004, online a http://digital.cs.usu.edu/~cyan/CS7960/hmm-tuto rial.pdf.

BOCCAFURNI - SERROMANI
1982 *Educazione linguistica nella scuola superiore: sei argomenti per un curricolo*, a cura di Anna Maria Boccafurni e Simonetta Serromani, Roma, Provincia di Roma - CNR, 1982.

BORGHETTI - CASTAGNOLI - BRUNELL
2011 Claudia Borghetti, Sara Castagnoli, Marco BrunellO, *I testi del web: una proposta di classificazione sulla base del corpus PAISÀ*, in CERRUTI - CORINO - ONESTI 2011, pp. 147-170.

BORGHI - RIZZA
2011 *Anatolistica Indoeuropeistica e Oltre – nelle Memorie dei Seminarî offerti da Onofrio Carruba (Anni 1997-2002), al Medesimo presentate*, a cura di Guido Borghi ed Alfredo Rizza, Milano, Qu.A.S.A.R, 2011 "Antiqui Aevi grammaticae artis studiorum consensus. Series maior" 1.

BOWKER - PEARSON
2002 Lynne Bowker, Jennifer Pearson, *Working with Specialized Languages. A Practical Guide to Using Corpora*, London - New York, Routledge, 2002.

BURNARD - BAUMAN
2011/08 *TEI P5:Guidelines for Electronic Text Encoding and Interchange*, edited by Lou Burnard and Syd Bauman, Charlottesville (Virginia), Text Encoding Initiative Consortium, 2011 [2008$_1$]; online alla pagina http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TitlePageVerso.html.

BURR
2004 Elisabeth Burr, *Das Korpus romanischer Zeitungssprachen in Forschung und Lehre*, in DAHMEN *et alii* 2004, pp. 133-62.
2005 *Tradizione & Innovazione.* [Volume I.] *Il parlato: teoria - corpora - linguistica dei corpora. Atti del VI Convegno SILFI (Gerhard Mercator Universität Duisburg 28 giugno - 2 luglio 2000)*, a cura di Elisabeth Burr, Firenze, Franco Cesati Editore, 2005 "Quaderni della rassegna" 43.

BUSA
1951 Roberto Busa SJ, *S. Thomae Aquinatis hymnorum ritualium varia specimina concordantiarum. Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate*, Milano, Bocca, 1951.
2005 Roberto Busa SJ, *Index Thomisticus, web edition by Eduardo Bernot and Enrique Alarcón*, 2005, online a http://www.corpusthomisticum.org/it/index.age

BUZZETTI
1999 Dino Buzzetti, *Rappresentazione digitale e modello del testo*, in AA. VV. 1999, pp. 127-161.

CARNAP
1937/34 Rudolf Carnap, *The Logical Syntax of Language*, English translation by Amethe Smeaton Countess von Zeppelin, London, Routledge & Kegan Paul, 1937. Edizione originale *Logische Syntax der Sprache*, Wien, 1934.
1974/63 Rudolf Carnap, *Autobiografia intellettuale*, in SCHILPP 1974, pp. 1-85 e 997-998. Edizione originale *Intellectual Autobiography*, in SCHILPP 1963.

CARPENTER
1992 Bob Carpenter, *The Logic of Typed Feature Structures. With Application to Unification Grammars, Logic Programs, and Constraint Resolution*, Cambridge (UK), Cambridge University Press, 1992 "Cambridge Tracts in Theoretical Computer Science" 32.

CERBO - DI FIORE
2011 *Lectura Dantis in onore di Vincenzo Placella*, a cura di Anna Cerbo e Ciro di Fiore, Napoli, Liguori, 2011.

CERRUTI - CORINO - ONESTI
2011 *Formale e informale. La variazione di registro nella comunicazione elettronica*, a cura di Massimo Cerruti, Elisa Corino e Cristina Onesti, Roma, Carocci Editore, 2011 "Biblioteca di testi e studi" 683.

CHOMSKY
1957/70 Noam Chomsky, *Syntactic Structure*, The Hague, Mouton, 1957. Versione italiana: *Le strutture della sintassi*, introduzione [traduzione e note] di Francesco Antinucci, Roma - Bari, Laterza, 1970 "Universale Laterza" 129.
1962/58 Noam Chomsky, *A Transformational Approach to Syntax*, presentato alla *3rd Texas Conference on Problems of Linguistic Analysis in English. May 9-12 1958*, poi in HILL 1962, pp. 124-158. CVfr. anche: *Discussion*.

1966/2002 Noam Chomsky, *Cartesian Linguistics. A Chapter in the History of Rationalist Thought*, New York, Harper & Row, 1966; ristampa: Lanham (MD) - New York (NY) - London (EN), University Press of America, 1983. Poi anche *Second Edition, edited with a new introduction* by James McGilvray, Christchurch (NZ), Cybereditions Corporation, 2002.

CHRIST - SCHULZE
1996 Oliver Christ - Bruno Maximilian Schulze, *CWB. Corpus Work Bench, Ein flexibles und modulares Anfragesystem für Textcorpora*, in FELDWEG 1996; disponibile online: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/Papers/christ+schulze:tuebingen.94.ps.gz.

CINI - REGIS
2002 *Che cosa ne pensa oggi Chiaffredo Roux? Percorsi di dialettologia percezionale all'alba del nuovo millennio. Atti del Convegno internazionale (Bardonecchia, 25-27 maggio 2000)*, a cura di Monica Cini e Riccardo Regis, Alessandria, Edizioni dell'Orso, 2002.

CIURCINA - RICOLFI
2007 Marco Ciurcina, Marco Ricolfi, *Le Creative Commons Public Licences per i corpora. Una suite di modelli per la linguistica dei corpora*, in BARBERA - CORINO - ONESTI 2007a, ¶ 7 pp. 127-132.

COOK - SEIDLHOFER
1995 *Principle & Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson,* edited by Guy Cook and Barbara Seidlhofer, Oxford - New York, Oxford University Press, 1995.

CORINO - MARELLO
2009a Elisa Corino - Carla Marello, *Elicitare scritti a partire da storie disegnate: il corpus di apprendenti Valico*, in ANDORNO - RASTELLI (2009), pp. 113-138.

2009b *VALICO. Studi di linguistica e didattica*, a cura di Elisa Corino e Carla Marello, Perugia, Guerra Edizioni, 2009.

2009c Elisa Corino - Carla Marello, *Didattica con i corpora di italiano per stranieri*, in «Italiano LinguaDue» I (2009) 279-285.

CORINO - MARELLO - ONESTI

2006 *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6-9 settembre 2006 | Proceedings of the XII EURALEX International Congress. Torino, Italia, 6th-9th September 2006*, a cura di Elisa Corino, Carla Marello e Cristina Onesti, Alessandria, Edizioni dell'Orso, 2 volumi, 2006.

COTTICELLI KURRAS

2011 *Linguistica e filologia digitale: aspetti e progetti*, a cura di Paola Cotticelli Kurras, Alessandria, Edizioni dell'Orso, 2011.

CRESTI- MONEGLIA

2005 *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*, edited by Emanuela Cresti and Massimo Moneglia, Amsterdam - Philadelphia, John Benjamins Publishing Company, 2005 "Studies in Corpus Linguistics" 15.

CREVATIN

2009 Franco Crevatin, *Al lettore*, in BARBERA 2009, p. iij.

DAHMEN *et alii*

2004 *Romanistik und neue Medien. Romanistisches Kolloquim XVI*, herausgegeben von Wolfgang Dahmen, Günter Holtus, Johannes Kramer, Michael Metzeltin, Wolfgang Schweickard und Otto Winkelmann, Tübingen, Günther Narr Verlag, 2004 "TBL" 4555.

DARRELL - TOMPA -WOOD

1992 R. Raymond Darrell - Frank W. Tompa - Derick Wood, *Markup Reconsidered*, paper presented at the *First International Workshop on Principles of Document Processing,*

Washington DC, October 22-23, 1992, 1992; disponibile: http:// db.uwaterloo.ca/~drraymon/papers/markup.ps.

DE HAAN
1992 Pietre De Haan, *The Optimum Corpus Sample Size?*, in LEITNER 1992, pp. 3-19.

DE MAURO - VOGHERA
1996 Tullio De Mauro - Miriam Voghera, *Scala mobile. Un punto di vista sui lessemi complessi*, in BENINCÀ *et alii* 1996, pp. 99-131.

DOMOKOS - SALVI
2002 *Lingue romanze nel Medioevo. Atti del convegno, Pilicsaba, 22-23 marzo 2002*, a cura di Domokos György e Salvi Giampaolo, in «Verbum. Analecta Neolatina» IV (2002)[2] 267-526.

EVERT *et alii*
2010a Stefan Evert - OCWB Development Team, *The IMS Open Corpus Workbench (CWB). Corpus Encoding Tutorial. CWB Version 3.0*, 2010; online alla pagina http://cwb.sourceforge.net/documentation.php.
2010b Stefan Evert - OCWB Development Team, *The IMS Open Corpus Workbench (CWB). CQP Query Language Tutorial. CWB Version 3.0*, 2010; online alla pagina http://cwb.sourceforge.net/documentation.php.

FALBO
2012 Caterina Falbo, *CorIT (Italian Television Corpus): classification criteria*, in STRANIERO - FALBO (2012), pp. 155-185.

FELDWEG - HINRICHS
1996 *Lexikon und Text: wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, herausgegeben von Helmut Feldweg, Erhard W. Hinrichs, Tübingen, Max Niemeyer Verlag, 1996 "Lexicographica. Series maior" 73.

FERRARI - LALA

2011 *Variétés syntaxiques dans la variété des textes online en italien: aspects micro- et macrostructuraux*, édité par Angela Ferrari, Letizia Lala, Nancy, Université de Nancy II, 2011 = «Verbum» XXXIII (2011)[1-2].

FILLMORE

1992 Charles J. Fillmore, *"Corpus Linguistics" or "Computer-aided Armchair Linguistics"*, in SVARTVIK 1992, pp. 35-60.

FRANCIS

1964 W[inthrop] N[elson] Francis, *A standard sample of present-day English for use with digital computers. Report to the US Office on Education on Co-operative Research Project no E-007*, Providence (Rhode Island), Brown University, 1964.

FRANCIS - KUČERA

1979/64 W[inthrop] N[elson] Francis - Henry Kucera [*sic*], *Brown Corpus Manual. Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*, Providence (Rhode Island), Brown University (Department of Linguistics), 1964; revised 1971; revised and amplified 1979.

FRIEDL

2006/1997 Jeffrey E. F. Friedl, *Mastering Regular Expression.* Third edition, O'Reilly Media, Beijing - Cambridge - Farnham - Köln - Sebastopol - Taipei - Tokyo, 2006 [2002_2, 1997_1].

FRIES C

1952 Charles Carpenter Fries, *The Structure of English; an Introduction to the Construction of English Sentences*, New York, Harcourt & Brace, 1952.

FRIES P

2010 Peter H. Fries, *Charles C. Fries, linguistics and corpus linguistics*, in «ICAME Journal» XXXIV (2010) 88-119.

GANDIN
2009 Stefania Gandin, *Linguistica dei corpora e traduzione: definizioni, criteri di compilazione e implicazioni di ricerca dei corpora paralleli*, in «AnnalSS» V (2005 [ma 2009]) 133-152.

GARSIDE - LEECH - MCENERY
1997 *Corpus Annotation. Linguistic Information from Computer Text Corpora*, edited by Roger Garside, Geoffrey Leech, Anthony McEnery, London - New York, Longman, 1997.

GARSIDE - SMITH
1997 Roger Garside - Nicholas Smith, *A Hybrid Grammatical Tagger: CLAWS4*, in GARSIDE - LEECH - MCENERY 1997, pp. 102-121.

GHADESSY - HENRY - ROSEBERRY
2002 *Small Corpus Studies and ELT: Theory and Practice*, edited by Mohsen Ghadessy, Alex Henry, Robert L. Roseberry, Amsterdam - Philadelphia, John Benjamins Pub Co., 2002 "Studies in Corpus Linguistics"

GRAFFI
1991 Giorgio Graffi, *Concetti 'ingenui' e concetti 'teorici' in sintassi*, in «Lingua e stile» XXVI (1991) 347-363.

GOOD
2004 Nathan A. Good, *Regular Expression Recipes: A Problem-Solution Approach*, Berkeley (CA), Apress, 2004.

GREFENSTETTE - TAPANAINEN
1994 Gregory Grefenstette - Pasi Tapanainen, *What is a Word, What is a Sentence? Problems of Tokenization*, in AA. VV. 1994, pp. 79-87, disponibile online alla pagina http://www.ling.helsinki.fi/~tapanain/tekeleet.html.

HÉDIARD
2007 *Linguistica dei corpora. Strumenti e applicazioni*, a cura di Marie Hédiard, Cassino, Edizioni dell'Università degli studi di Cassino, 2007 "Collana Scientifica" 20.

HEID

1998 Ulrich Heid, *Annotazione morfosintattica di corpora ed estrazione di informazioni linguistiche*, relazione al convegno *Annotazione morfosintattica di corpora e costruzione di banche di dati linguistici. Torino, 26-XI-1998*, inedito,1998.

2007 Ulrich Heid, *Il* Corpus WorkBench *come strumento per la linguistica dei corpora. Principi ed applicazioni*, in BARBERA - CORINO - ONESTI 2007a, ¶ 4 pp. 89-108.

HILL

1962 *Third Texas Conference on Problems of Linguistic Analysis in English: May 9-12, 1958*, edited by A[rchibald] A. Hill, Austin, University of Texas, 1962 "Studies in American English".

IANNACCARO

i.s. *La linguistica italiana all'alba del terzo millennio (1997-2010)*, a cura di Gabriele Iannaccaro Roma, Bulzoni, in corso di stampa "SLI Società di linguistica italiana".

JAFRACESCO

2011 *L'acquisizione del lessico nell'apprendimento dell'italiano L2. Atti del XIX convegno nazionale ILSA, Firenze, 27 novembre 2010*, a cura di Elisabetta Jafracesco, Firenze, Le Monnier, 2011 "Italiano per stranieri".

JURAFSKY - MARTIN

2000 Daniel [Saul] Jurafsky - James H. Martin, *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Upper Saddle River (NJ), Prentice Hall, 2000 "Prentice Hall Series in Artificial Intelligence".

KETTEMANN - MARKO

2000 *Language and Computers, Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora,*

Graz 19-24 July, 2000, edited by Bernhard Kettemann and Georg Marko, Amsterdam, Rodopi, 2000.

KILGARRIFF - GREFENSTETTE
2003 Adam Kilgarriff - Gregory Grefenstette, *Introduction to the Special Issue on the Web as Corpus*, in «Computational Linguistics» XXIX (2003)[3] 333-347, disponibile anche online alla pagina http://www.kilgarriff.co.uk/publications.htm.

KLAVANS - RESNIK
1996 *The Balancing Act. Combining Symbolic and Statistical Approaches to Language*, edited by Judith L. Klavans and Philip Resnik, Cambridge (Mass.) - London (GB), MIT Press, 1996.

LEECH
1991 Geoffrey Leech, *The State or the Art in Corpus Linguistics*, in AIJMER - ALTENBERG 1991, pp. 8-29.

LEITNER
1992 *New Directions in English Language Corpora: Methodology, Results, Software Developments*, edited by Gerhard Leitner, Berlin, Mouton de Gruyter, 1992 “Topics in English Linguistics”.

LEMNITZER - ZINSMEISTER
2006 Lothar Lemnitzer, Heike Zinsmeister, *Korpuslinguistik: eine Einführung*, Tübingen, Gunter Narr Verlag, 2006 “Narr Studienbücher”.

LENCI - MONTEMAGNI - PIRRELLI
2005 Alessandro Lenci - Simonetta Montemagni - Vito Pirrelli, *Testo e computer. Elementi di linguistica computazionale*, Roma, Carocci, 2005 “Università” 664.

LEOPARDI
1991/1817-27 Giacomo Leopardi, *Zibaldone di pensieri*, edizione critica e annotata a cura di Giuseppe Pacella, Milano, Garzanti, 1991 “I libri della spiga”.

LESMO - LOMBARDO - BOSCO
2002 Leonardo Lesmo - Vincenzo Lombardo - Cristina Bosco, *Treebank Development: the TUT Approach*, in AA. VV. 2002; online a http://www.di.unito.it/~tutreeb/.

LEWANDOWSKA-TOMASZCZYK - OSBORNE - SCHULTE
2001 Barbara Lewandowska-Tomaszczyk - John Osborne - Frits Schulte, *Foreign Language Teaching and Information and Communication Technology*, Frankfurt am Main, Peter Lang, 2001 "Łódź Studies in Language" 3.

LÜDELING - KYTÖ
2008-9 *Corpus Linguistics, An International Handbook*, edited by Anke Lüdeling and Merja Kytö, Berlin Mouton de Gruyter, vol. I 2008, vol. II 2009 "Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science" 29.1-2.

LUHN
1960 H[ans] P[eter] Luhn, *Keyword-in-context Index for Technical Literature (KWIC Index)*, in «America Documentation» XI (1960) 288-295.

LYDING *et alii*
2006 Verena Lyding - Elena Chiocchetti - Gilles Sérasset - Francis Brunet-Manquat, *The LexALP Information System: Term Bank and Corpus for Multilingual Legal Terminology Consolidated,* in AA.VV. 2006a, pp. 25–31.

MACWHINNEY
2000 Brian MacWhinney, *The CHILDES Project: Tools for Analyzing Talk*. Volume 1: *Transcription format and programs*. Volume 2: *The Database*, Mahwah (NJ), Lawrence Erlbaum Associates, 2000.

MANNING - SCHÜTZE
1999 Christopher D. Manning - Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, Cambridge (Massachusetts) - London (England), The MIT Press, 1999.

MANZONI
1975 Giacomo Manzoni, *Arnold Schönberg, L'uomo, l'opera, i testi musicati. Con una composizione giovanile inedita*, Milano, Feltrinelli Editore, 1975 "Universale Economica" 725.

MARASCHIO - DE MARTINO - STANCHINA
2011 *L'italiano degli altri. Atti (Firenze, 27-31 maggio 2010)*, a cura di Nicoletta Maraschio, Domenico De Martino e Giulia Stanchina, Firenze, Accademia della crusca, 2011 "La Piazza delle lingue" 2.

MARCUS - SANTORINI - MARCINKIEVICZ
1984 Mitchell P. Marcus - Beatrice Santorini - Marcinkievicz Mary Ann Marcinkievicz, *Building a Large Annotated Corpus of English: The Penn Treebank*, in ARMSTRONG 1994, pp. 273-290. Disponibile online dalla homepage del PennTreebank al link ftp://ftp.cis.upenn.edu/pub/treebank/doc/cl93.ps.gz.

MARCKWARDT
1968 Albert H. Marckwardt, *Charles C. Fries*, in «Language», XLIV (1968)[1] 205-210.

MARELLO
1996 Carla Marello, *Le parole dell'italiano. Lessico e dizionari,* Bologna, Zanichelli, 1996.

MCENERY - HARDIE
2012 Tony McEnery - Andrew Hardie, *Corpus Linguistics: Method, Theory and Practice*, Cambridge, Cambridge University Press, 2012 "Cambridge textbooks in linguistics".

MCENERY - WILSON
2001 Tony McEnery - Andrew Wilsom (2001/1996), *Corpus Linguistics. An Introduction*, Edinburgh, Edinburgh University Press, 2001 [1996_1] "Edinburgh Textbooks in Empirical Linguistics".

MEYER
2002 Charles F. Meyer (2002), *English Corpus Linguistics. An Introduction*, Cambridge, Cambridge University Press, 2002 “Studies in English Languages”.

MITKOV
2003 *The Oxford Handbook of Computational Linguistics*, edited by Ruslan Mitkov, Oxford, Oxford University Press, 2003.

MONACHINI
1996 Monica Monachini, *ELM-IT: EAGLES Specifications for Italian Morphosyntax - Lexicon Specifications and Classification Guidelines*, Pisa, 1996 EAGLES Document EAG-CLWG-ELM-IT/F. Disponibile online alla pagina: http://www.ilc.cnr.it/EAGLES/browse.html.

MONACHINI - CALZOLARI
1996 Monica Monachini - Nicoletta Calzolari, S*ynopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Application to European Languages*, Pisa, 1996, EAGLES Document EAG-CLWG-MORPHSYN/R, May. Disponibile online alla pagina: http://www.ilc.cnr.it/EAGLES/browse.html.

MONTEMAGNI *et alii*
2003 Simonetta Montemagni - Francesco Barsotti - Marco Battista - Nicoletta Calzolari - Ornella Corazzari - Alessandro Lenci - Antonio Zampolli - Francesca Fanciulli - Maria Massetani - Remo Raffaelli - Roberto Basili - Maria Teresa Pazienza - Dario Saracino - Fabio Zanzotto - Nadia Mana - Fabio Pianesi - Rodolfo Delmonte, *Building the Italian Syntactic-Semantic Treebank*, in ABEILLÉ 2003, pp. 189-210; disponibile online a http://www.ilc.cnr.it/tressi_prg/papers/isst_kluwer_final_version.pdf.

MORALDO
2008 *Sprachkontakt und Mehrsprachigkeit. Zur Anglizismendiskussion in Deutschland, Österreich, der Schweitz und Ita-*

lien. Convegno Iternazionale, Forlì, 21-22 marzo 2007, herausgegeben von Sandro M. Moraldo, Heidelberg, Universitätsverlag Winter, 2008.

MORTARA GARAVELLI
2001 Bice Mortara Garavelli, *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*, Torino, Giulio Einaudi Editore, 2001 "Piccola biblioteca Einaudi" 150.

OAKES
1998 Michael P. Oakes, *Statistics for Corpus Linguistics*, Edinburgh, Edinburgh University Press, 1998 "Edinburgh Textbooks in Empirical Linguistics".

NESI - DE MARTINO
2012 *Lingua italiana e scienze. Atti del convegno internazionale, Firenze, Villa Medicea di Castello, 6-8 febbraio 2003* a cura di Annalisa Nesi e Domenico De Martino, Firenze, Accademia della Crusca, 2012.

ONESTI
2011 Cristina Onesti, *Methodology for Building a Text-Structure Oriented Legal Corpus*, in «Comparative Legilinguistics» VIII (2011) 37-50.

PAJUSALU - HENNOSTE
2002 *Tähendusepüüdja. Pühendusteos professor Haldur Õimu 60. sünnipäevaks | Catcher of the Meaning. A Festschrift for Professor Haldur Õim*, toimetanud Renate Pajusalu ja Tiit Hennoste, Tartu, Tartu Ülikooli Kirjastus, 2002 "Tartu Ülikooli üldkeeleteaduse õppetooli toimetised | Publications of the Department of General Linguistics" 3.

PALERMO
2009 *Percorsi e strategie di apprendimento dell'italiano lingua seconda: sondaggi su ADIL2*, a cura di Massimo Palermo, Perugia, Guerra Edizioni, 2009 "Università per stranieri di Siena. Materiali del centro di eccellenza" 5.

PANDOLFI
2006 Elena Maria Pandolfi, *Misurare la regionalità. Uno studio quantitativo su regionalismi e forestierismi nell'italiano parlato nel Canton Ticino*, Locarno, Osservatorio linguistico della Svizzera italiana - Armando Dadò, 2006.

PEIRCE
1906/2006 Charles Sanders Peirce, *Prolegomena to an Apology for Pragmaticism*, in «The Monist» XVI (1906)[4] 492-546; poi in *Collected Papers of Charles Sanders Peirce*, 8 volumes (vols. 1-6, eds. Charles Hartshorne and Paul Weiss, vols. 7-8, ed. Arthur W. Burks), Cambridge (Mass.), Harvard University Press, 1931-1958, vol. IV. Traduzione italiana prima in PEIRCE 1980, pp. 211-271 e poi in PEIRCE 2011, pp. 205-250.
1980 Charles Sanders Peirce, *Semiotica*, testi scelti e introdotti da Massimo A. Bonfantini, Letizia Grassi, Roberto Grazia, Torino, Einaudi , 1980 "Einaudi Paperbacks e Readers".
2011 Charles Sanders Peirce, *Opere*, a cura di Massimo Bonfantini, con la collaborazione di Giampaolo Proni, Milano, Bompiani, 2011 [2003_1] "Il pensiero occidentale".

POWELL
2006 Christophers Powell, *SEMiSUSANNE Corpus: Documentation*, 2006, pagina web http://www.grsampson.net/Semi SueDoc.html. [This Web version of Christopher Powell's SEMiSUSANNE readme file was prepared by Geoffrey Sampson on 17 Jan 2006].

QUINE
1987 Willard van Orman Quine, *Quiddities: an Intermittently Philosophical Dictionary*, Cambridge (Mass.), the Belknap Press of Harvard University Press, 1987.

RABINER
1989 Lawrence R. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, in «Proceedings of IEEE» LXXVII (1989)[2], pp. 257-286.

RENZI

2002 Lorenzo Renzi, *Il progetto ItalAnt e la "grammatica del corpus"*, in DOMOKOS - SALVI 2002, pp. 271-94.

2008/02 Lorenzo Renzi, *L'autobiografia linguistica in generale, e quella dell'autore in particolare, con un saggio di quest'ultima*, in CINI - REGIS 2002, pp. 329-339, poi in RENZI 2008, pp. 3-16.

2008 Lorenzo Renzi, *Le piccole strutture. Linguistica, poetica, letteratura*, a cura di Alvise Andreose, Alvaro Barbieri, Dan Octavian Cepraga, con la collaborazione di Marina Doni, Società editrice il Mulino, Bologna.

RENZI - SALVI

2010 *Grammatica dell'italiano antico*, a cura di Lorenzo Renzi e Giampaolo Salvi, 2 volumi, Bologna, il Mulino, 2010.

ROBBINS

2012 Arnold D. Robbins, *GAWK: Effective AWK Programming: A User's Guide for GNU AWK*, edition 4, Boston, Free Software Foundation, 2012; online alla pagina: http://www.gnu.org/software/gawk/manual/.

ROSSINI FAVRETTI

1998 Rema Rossini Favretti, *Using Multilingual Parallel Corpora for the Analysis of Legal Language: the Bononia Legal Corpus*, in TEUBERT - TOGNINI-BONELLI - VOLZ (1998), pp. 57-68; disponinile online: http://corpora.dslo.unibo.it/BoLCPubs/Rossini1998_UsingMultilingual.pdf.

2000a *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*, a cura di Rema Rossini Favretti, Roma, Bulzoni, 2000.

2000b Rema Rossini Favretti, *Progettazione e costruzione di un corpus di italiano scritto: CORIS/CODIS*, in ROSSINI FAVRETTI 2000, pp. 39-56; disponiblie online: http://corpora.dslo.unibo.it/CORISPubs/Rossini2000_ProgettazioneeCostruzione.pdf.

ROVERE
2005 Giovanni Rovere, *Capitoli di linguistica giuridica. Ricerche su corpora elettronici*, Alessandria, Edizioni dell'Orso, 2005 "Gli argomenti umani" 9.

SABATINI
2011/1982 Francesco Sabatini, *La comunicazione orale, scritta e trasmessa: la diversità del mezzo, della lingua e delle funzioni*, in SABATINI 2011, vol. II, pp. 55-77; già in BOCCAFURNI - SERROMANI 1982, pp. 105-127.
2011/06 Francesco Sabatini, *La storia dell'italiano nella prospettiva della* corpus linguistics, in SABATINI 2011, vol. I, pp. 223-232; già in CORINO - MARELLO - ONESTI 2006, pp. 31-37.
2007 Francesco Sabatini, *Storia della lingua italiana e grandi corpora. Un capitolo di storia della linguistica*, in BARBERA - CORINO - ONESTI 2007a, ¶ ij pp. xiij-xvj.
2011/08 Francesco Sabatini, *L'italiano lingua permissiva? Proposta per una strategia comune delle lingue europee verso l'anglicismo*, in SABATINI 2011, vol. III, pp. 333-341; già in MORALDO 2008, pp. 267-275.
2011 Francesco Sabatini, *L'italiano nel mondo moderno. Storia degli usi e della norma, la scuola, i dialetti, il latino, modelli teorici, la Crusca, l'Europa. Saggi dal 1968 al 2009*, a cura di Vittorio Coletti, Rosario Coluccia, Paolo d'Achille, Nicola de Blasi, Domenico Proietti, Napoli, Liguori editore, 2011, 3 volumi.

SAMPSON
1997 Geoffrey Sampson, *Educating Eve. The 'Language Instinct' Debate*, London - New York, Cassel, 1997 "Open Linguistics".
2001 Geoffrey Sampson, *Empirical Linguistics*, London - New York, Continuum, 2001 "Open Linguistics".
2004 Geoffrey Sampson, [*Foreword*] to *Charles Carpenter Fries, from The Structure of English. 1952*, in SAMPSON - MCCARTHY 2004, p. 9.

SAMPSON - MCCARTHY
2004 *Corpus Linguistics. Readings in a Widening Discipline*, edited by Geoffrey Sampson and Diana McCarthy, London - New York, Continuum, 2004.

SAUSSURE
2001/1916 Ferdinand de Saussure, *Cours de linguistique générale*, publié par Charles Bailly et Albert Séchehaye, avec la collaboration de Albert Riedingler, édition critique préparée par Tullio de Mauro, postface de Louis-Jean Calvet, Paris, Payot, 2001 “Grande bibliothèque Payot”. Edizione originaria: *ibidem*, 1916. Edizione italiana: *Corso di linguistica generale*, introduzione, traduzione e commento di Tullio De Mauro, Roma - Bari, Laterza, 1967_1.

SCHMID
1994 Helmut Schmid, *Probabilistic Part-of-Speech Tagging Using Decision Trees*, paper presented at the *International Conference on New Methods in Language Processing, Manchester (UK), 1994*; versione revisionata PS/PDF online sul sito dell’IMS Stuttgart: http://www.ims.uni-stuttgart.de/projekte/corplex/Tree Tagger/DecisionTreeTagger.html.

SCHILPP
1974/63 *La filosofia di Rudolf Carnap*, a cura di Paul Arthur Schilpp, traduzione di Maria Grazia Cristofaro Sandrini, Milano, il Saggiatore, 1974 “Biblioteca di filosofia e metodo scientifico”, pp. 1-85 e 997-998. Edizione originale *The Philosophy of Rudolf Carnap*, edited by P[aul] A[rthur] Schilpp, La Salle (Illinois), 1963 “The Library of Living Philosophers”.

SCHÖNBERG
1950/33 Arnold Schönberg, *Brahms il progressivo*, conferenza tenuta il 12 febbraio 1933 e poi riprodotta rielaborata in SCHÖNBERG 1950/60, pp. 56-104.
1950/60 Arnold Schönberg, *Style and Idea*, New York, Philosophical Library, 1950. Traduzione italiana di Maria Giovanna Moretti e Luigi Pestalozza: *Stile e idea*, con un saggio di

Luigi Pestalozza, Milano, Feltrinelli, 1980_3 [1975_2, 1960_1] “I fatti e le idee. Saggi e biografie” 293.

SIMONE
2010-11 *Enciclopedia dell’italiano*, a cura di Raffaele Simone, con la collaborazione di Gaetano Berruto e Paolo D’Achille, Roma, Istituto dell’Enciclopedia italiana fondata da Giovanni Treccani, vol. I. 2010 e vol. II. 2011 “Vocabolario Treccani”.

SINCLAIR
1991 John [McHardy] Sinclair, *Corpus, Concordance, Collocation*, Oxford, Oxford University Press, 1991.

SKINNER
1957 Frederik B[urrhus] Skinner, *Verbal Behaviour*, New York, Appleton - Century - Crofts = London, Methuen, 1957.

SPINA
2005/00 Stefania Spina, *Il Corpus di italiano televisivo (CiT): struttura e annotazione*, in BURR 2005, pp. 413-426.

STEFANELLI - MARASCHIO
2003 *LIR - Lessico Italiano Radiofonico (1995-2003)*, a cura di Stefania Stefanelli e Nicoletta Maraschio, Firenze, Accademia della Crusca, 2003, 2 DVD.

STOPPELLI - PICCHI
2001 *LIZ 4.0. Letteratura italiana Zanichelli. CD-ROM dei testi della letteratura italiana*, a cura di Pasquale Stoppelli e Eugenio Picchi, Bologna, Zanichelli, 2001 quarta edizione per Windows.

STRANIERO
2007 Sergio Francesco Straniero, *Talkshow interpreting. La mediazione linguistica nella conversazione-spettacolo*, Trieste, EUT - Edizioni Università di Trieste, 2007.

STRANIERO - FALBO
2012 *Breaking Ground in Corpus-based Interpreting Studies*, edited by Sergio Francesco Straniero and Caterina Falbo, Bern - Berlin - Bruxelles - Frankfurt am Main - New York

- Oxford - Wien, Peter Lang, 2012 “Linguistic Insights” 147.

SVARTVIK
1992 *Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82. Stockholm, 4-8 August 1991*, edited by Jan Svartvik, Berlin, Mouton de Gruyter, 1992 “Trends in Linguistics. Studies and Monographs” 65.

SWAFFORD
1998 Jan Swafford, *Johannes Brahms. A Biography*, New York, Alfred A. Knopf, 1998.

TAVONI
2011 Mirko Tavoni, *DanteSearch: il corpus delle opere volgari e latine di Dante lemmatizzate con marcatura grammaticale e sintattica*, in CERBO - DI FIORE 2011, pp. 567-591.
2012 *Sintassi dell’Italiano antico e sintassi di Dante*, a cura di Mirko Tavoni, Pisa, Felici Editore, 2012. Basato sulle comunicazioni al *Seminario: sintassi dell’italiano antico e sintassi di Dante. Pisa 14-15 ottobre 2011*.

TEUBERT - TOGNINI-BONELLI - VOLZ
1998 *Translation Equivalence. Proceedings of the Third European Seminar*, edited by Wolfgang Teubert, Elena Tognini-Bonelli and Norbert Volz, Mannheim, The TELRI Association - Institut für Deutsche Spräche - The Tuscan Word Centre, 1998.

TOGNINI-BONELLI
2001 Elena Tognini-Bonelli, *Corpus Linguistics at Work*, Amsterdam - Philadelphia, John Benjamins Publishing Company, 2001 “Studies in Corpus Linguistics” 6.

TOMATIS
2007 Marco [Stefano] Tomatis, *La disambiguazione del Corpus Taurinense. Problemi teorici e pratici*, in BARBERA - CORINO - ONESTI 2007a, ¶ 9 pp. 169-181.

TRIBBLE

1997 Chris Tribble, *Improvising Corpora for ELT: Quick and Dirty Ways of Developing Corpora for Language Teaching*, in LEWANDOWSKA-TOMASZCZYK - MELIA 1997. HTML online alla pagina http://www.ctribble.co.uk/text/Palc.htm.

VILLARINI

2008 Andrea Villarini, *Il lessico dei materiali didattici usati nei corsi di italiano per immigrati*, in BARNI - TRONCARELLI - BAGNA 2008, pp. 165-177.

2011 Andrea Villarini, *La competenza lessicale: un viaggio tra libri di testo e parlato del docente*, in JAFRACESCO 2011, pp. 53-80.

VOLK

2002 Martin Volk, *Using the Web as Corpus for Linguistic Research*, in PAJUSALU - HENNOSTE (2002). Disponibile online alla pagina http://www.ling.su.se/DaLi/volk/publications.html.

VOLTOLINI

1998/2002 Alberto Voltolini, *Internalism & Externalism. [Second Draft]*, online paper, 1998 last updated 2002: online a: http://host.uniroma3.it/progetti/kant/field/voltolini.html.

WATSON

1913 John Broadus Watson, *Psychology as a Behaviorist Views It*, in «Psychological Review» XX (1913) 158-77.

ZANETTIN

2000 Federico Zanettin, *CEXI: Designing an English Italian Translational Corpus*, in KETTEMANN - MARKO 2000, pp. 329-343

ZANNI

2007 Zanni Samantha, *Corpora elettronici e copyright. Lo stato legale della questione*, in BARBERA - CORINO - ONESTI (2007a), ¶ 6 pp. 119-126.

4.2.2 Corpora e siti riferiti.

ADAM
Dialoghi annotati per interfacce vocali avanzate
http://www.ilc.cnr.it/viewpage.php/sez=ricerca/id=871/vers=ita

Allora HP
http://www.e-allora.net/

ANC
American National Corpus
http://www.americannationalcorpus.org/

AntConc
http://www.antlab.sci.waseda.ac.jp/antconc_index.html

Athenaeum Corpus
http://www.bmanuel.org/projects/at-HOME.html

BADIP
BAnca Dati dell'Italiano Parlato
http://badip.uni-graz.at/

Barbera HP
http://www.bmanuel.org/personal/barbera/barbera.html

Baroni HP
http://clic.cimec.unitn.it/marco/index.html

bmanuel.org
http://www.bmanuel.org/

BoLC
BOnonia Legal Corpus
http://corpora.dslo.unibo.it/bolc_eng.html

BNC
British National Corpus
http://www.natcorp.ox.ac.uk/

Canterbury Corpus
Evaluating lossless compression methods
http://corpus.canterbury.ac.nz/

CC
Creative Commons
http://creativecommons.org/
http://it.creativecommons.org/

CEOD
Corpus Epistolare Ottocentesco Digitale
http://ceod.unistrasi.it/

CHILDES-it
CHIld Language Data Exchange System - Italian
http://childes.psy.cmu.edu/data/Romance/Italian/

CIBID
Centro Interuniversitario Biblioteca Italiana Digitale
http://www.bibliotecaitaliana.it/cibit/cibit.php

CLAN
Computerized Language ANalysis
http://childes.psy.cmu.edu/clan/

CLAWS
http://ucrel.lancs.ac.uk/claws/

CLIPS
Corpora e Lessici dell'Italiano Parlato e Scritto
http://www.clips.unina.it/it

ČNK
Český Národní Korpus | *Czech National Corpus*
http://ucnk.ff.cuni.cz/english/index.php

collocations.de
http://www.collocations.de/

C-ORAL ROM
Integrated reference corpora for spoken romance languages
http://lablita.dit.unifi.it/coralrom/intro.html
http://www.elda.org/catalogue/en/speech/S0172.htm

CORIS
CORpus di Italiano Scritto
http://corpora.dslo.unibo.it/coris_ita.html

Corpora list
http://www.hit.uib.no/corpora/

corpora.unito.it
http://www.corpora.unito.it

Corpus Segusinum
Corpus “La Valsusa”
http://www.bmanuel.org/projects/vs-HOME.html

Crusca online
Lessicografia della Crusca in Rete
http://www.lessicografia.it/

CT
Corpus Taurinense
http://www.bmanuel.org/projects/ct-HOME.html

CWB & CQP
Corpus WorkBench & Corpus Query Processor
http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/
http://cwb.sourceforge.net/

DanteSearch
http://dante.di.unipi.it:8080/DanteWeb/

DeReKo
Das Deutsche Referenzkorpus
http://www.ids-mannheim.de/kl/projekte/korpora/

Dia-LIT
Lessico DIAcronico dell’Italiano Televisivo
http://193.205.158.203/dialit/

EAGLES
Expert Advisory Group on Language Engineering Standards
http://www.ilc.cnr.it/EAGLES96/home.html

ELDA
Evaluations and Language resources Distribution Agency
http://www.elda.org/

EPIC
European Parliament Interpreting Corpus
http://dev.sslmit.unibo.it/corpora/corporaproject.php?path=E.P.I.C.

EURAC
EURopean ACademy of Bozen/Bolzano
http://www.eurac.edu/it/eurac/welcome/default.html

Evert HP
http://www.stefan-evert.de/

FSF
Free Software Fundation
www.fsf.org

GAWK
http://www.gnu.org/software/gawk/gawk.html

Google Libri
http://books.google.it/bkshp?hl=it&tab=wp

HNC
Hellenic National Corpus | Σώμα κειμένων του Ινστιτούτου Επεξεργασίας του Λόγου
http://hnc.ilsp.gr/default.asp

HNK
Hrvatski Nacionalni Korpus
http://www.hnk.ffzg.hr/

ILC
Istituto di Linguistica Computazionale "Antonio Zampolli"
http://www.ilc.cnr.it/indexflash.html

IMS Stuttgart
Institut für Maschinelle Sprachverarbeitung - Universität Stuttgart
http://www.ims.uni-stuttgart.de

Index Thomisticus
Corpus Thomisticum - Index Thomisticus
http://www.corpusthomisticum.org/it/index.age

InfoIus
http://www.infoius.it/nuovo_forum/default.asp

InfoLeges
http://www.infoleges.it/

ISST
Italian Syntactic-Semantic Treebank
http://medialab.di.unipi.it/isst/ISST.html

Italiano televisivo
Il portale dell'italiano televisivo
http://www.italianotelevisivo.org/

itWaC
→ WaCky Corpora

IULA etiquetaris
Institut Universitari de Lingüística Aplicada - Etiquetaris
http://www.iula.upf.edu/corpus/etiqueca.htm

IULA corpus
Institut Universitari de Lingüística Aplicada - Corpus textual especialitzat plurilingüe
http://www.iula.upf.edu/corpus/corpus.htm

Juris Data
http://www.giuffreroma.it/juris_data.html

Jus Jurium
http://www.bmanuel.org/projects/ju-HOME.html

LABLITA
Laboratorio Linguistico del Dipartimento di Italianistica dell'Università di Firenze
http://lablita.dit.unifi.it

LAICO
Lessico per Apprendere l'Italiano. Corpus di Occorrenze
https://sites.google.com/site/corpuslaico/

LCCPW
Lancaster Corpus of Children's Project Writing
http://www.lancs.ac.uk/fass/projects/lever/index.htm

LexAlp
http://lexalp.eurac.edu/projects/corpus_it.htm
http://lexalpapps.eurac.edu:8080/htdocs2/lexalp/corp_lexalp/search_corp.php

LiberLiber
http://www.liberliber.it/home/index.php

LIP
Corpus del Lessico di frequenza dell'Italiano Parlato
http://badip.uni-graz.at/index.php?option=com_content&view=article&id=8&Itemid=8&lang=it

LIT
Lessico dell'Italiano Televisivo
http://deckard.micc.unifi.it:8080/litsearch/
http://193.205.158.203/lit_ric2/

LLC
London-Lund Corpus of Spoken English
http://www.helsinki.fi/varieng/CoRD/corpora/LLC/index.html

METER Corpus
Journalistic Text Reuse Corpus
http://www.dcs.shef.ac.uk/nlp/meter/index.html

MNSz
Magyar Nemzeti Szövegtár | Hungarian National Corpus
http://corpus.nytud.hu/mnsz/

NKJP
Narodowy Korpus Języka Polskiego | National Corpus of Polish
http://nkjp.pl/index.php

NKRJa
Национальный Корпус Русского Языка | Russian National Corpus
http://ruscorpora.ru/index.html

NUNC
Newsgroups UseNet Corpora
http://www.bmanuel.org/projects/ng-HOME.html

OVI
Opera del Vocabolario Italiano
http://www.vocabolario.org/

OVI banca dati
Corpus TLIO
http://tlioweb.ovi.cnr.it

PAISÀ
Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati
http://www.corpusitaliano.it/it/

Penn Treebank
The Penn Treebank Project
http://www.cis.upenn.edu/~treebank/home.html

Perl
http://www.perl.org/

PPCME
Penn-Helsinki Parsed Corpus of Middle English
http://www.ling.upenn.edu/histcorpora/PPCME2-RELEASE-3/index.html

Prague Treebank
Pražský závislostní korpus | The Prague Dependency Treebank V. 2.0
http://ufal.mff.cuni.cz/pdt2.0/

R
The R Project for Statistical Computing
http://www.r-project.org/

ReC
“la Repubblica” Corpus
http://dev.sslmit.unibo.it/corpora/corpus.php?path=&name=Repubblica

regexp_tokenizer
http://sslmit.unibo.it/~baroni/regexp_tokenizer.html

Sampson HP
http://www.grsampson.net/index.html

SCP
Simple Concordance Program
http://www.textworld.com/scp

SLI
Società di Linguistica Italiana
http://www.societadilinguisticaitaliana.net/

SMS Monitor Studies
http://www.e-allora.net/SMS/ms_index.php

Stein HP
http://www.uni-stuttgart.de/lingrom/stein/

SUSANNE
The SUSANNE Corpus & Analytic Scheme
http://www.grsampson.net/RSue.html

Tagset Baroni
http://sslmit.unibo.it/~baroni/collocazioni/itwac.tagset.txt

Tamburini HP
http://corpora.dslo.unibo.it/People/Tamburini/

TEI
Text Encoding Initiative
http://www.tei-c.org/

TLIO
Tesoro della Lingua Italiana delle Origini
http://tlio.ovi.cnr.it/

Tomatis HP
http://www.bmanuel.org/personal/tomatis/tomatis.html

TreeTagger
http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

TUT
Turin University Treebank
http://www.di.unito.it/~tutreeb/

VALICO
Varietà di Apprendimento della Lingua Italiana Corpus Online
http://www.valico.org/

VINCA
Varietà di Italiano di Nativi Corpus Appaiato
http://www.valico.org/vinca_CORPUS.html

VIT
Venice Italian Treebank
http://www.elda.org/catalogue/en/text/W0040.html

XML
Extensible Markup Language (XML)
http://xml.coverpages.org/xml.html

WaCky Corpora
Web as Corpus Kool Ynitiative corpora
http://wacky.sslmit.unibo.it/doku.php

WordLister
http://www.bmanuel.org/tools/WordLister/WordLister.html

Wordsmith's Tools
http://www.lexically.net/wordsmith/

zipfR
user-friendly LNRE modelling in R
http://zipfr.r-forge.r-project.org/

Indice generale.